Active Inference for Stochastic Control

Aswin Paul[1,2,3], Noor Sajid[4], Manoj Gopalkrishnan[2], and Adeel Razi[3,4,5,6]

[1] IITB-Monash Research Academy, Mumbai, India
[2] Department of Electrical Engineering, IIT Bombay, Mumbai, India
[3] Turner Institute for Brain and Mental Health, Monash University, Australia
[4] Wellcome Trust Centre for Human Neuroimaging, UCL, United Kingdom
[5] Monash Biomedical Imaging, Monash University, Australia
[6] CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, Canada

Abstract. Active inference has emerged as an alternative approach to control problems given its intuitive (probabilistic) formalism. However, despite its theoretical utility, computational implementations have largely been restricted to low-dimensional, deterministic settings. This paper highlights that this is a consequence of the inability to adequately model stochastic transition dynamics, particularly when an extensive policy (i.e., action trajectory) space must be evaluated during planning. Fortunately, recent advancements propose a modified planning algorithm for finite temporal horizons. We build upon this work to assess the utility of active inference for a stochastic control setting. For this, we simulate the classic windy grid-world task with additional complexities, namely: 1) environment stochasticity; 2) learning of transition dynamics; and 3) partial observability. Our results demonstrate the advantage of using active inference, compared to reinforcement learning, in both deterministic and stochastic settings.

Keywords: Active inference · Optimal control · Stochastic control · Sophisticated inference

1 Introduction

Active inference, a corollary of the free energy principle, is a formal way of describing the behaviour of self-organising systems that interface with the external world and maintain a consistent form over time [1,2,3]. Despite its roots in neuroscience, active inference has snowballed to many fields owing to its ambitious scope as a general theory of behaviour [4,5,6]. Optimal control is one such field, and several recent results place active inference as a promising optimal control algorithm [7,8,9]. However, research in the area has largely been restricted to low-dimensional and deterministic settings where defining, and evaluating, policies (i.e., action trajectories) is feasible [9]. This follows from the active inference process theory that necessitates equipping agents a priori with sequences of actions in time. For example, with 8 available actions and a time-horizon of 15, the total number of (definable) policies that would need to be considered $\rightarrow 3.5 \times 10^{13}$.

This becomes more of a challenge in stochastic environments with inherently uncertain transition dynamics, and no clear way to constrain the large policy space to a smaller subspace. Happily, recent advancements like sophisticated inference [10] propose a

modified planning approach for finite-temporal horizons [11]. Briefly, sophisticated inference [10], compared to the earlier formulation [12,9], provides a recursive form of the expected free energy that implements a deep tree search over actions (and outcomes) in the future. We reserve further details for Section 3.2.

In this paper, we evaluate the utility of active inference for stochastic control using the sophisticated planning objective. For this, we utilise the windy grid-world task [13], and assess our agent's performance when varying levels of complexity are introduced e.g., stochastic wind, partial observability, and learning the transition dynamics. Through these numerical simulations, we demonstrate that active inference, compared to a Q-learning agent [13], provides a promising approach for stochastic control.

2 Stochastic control in a windy grid-world

In this section, we describe the windy grid-world task, with additional complexity, used for evaluating our active inference agent (Section 3). This is a classic grid-world task from reinforcement learning [13], with a predefined start (S) and goal (G) states (Fig. 1). The aim is to navigate as optimally (i.e., within a minimum time horizon) as possible, taking into account the effect of the wind along the way. The wind runs upward through the middle of the grid, and the goal state is located in one such column. The strength of the wind is noted under each column in Fig. 1, and its amplitude is quantified by the number of columns shifted upwards that were unintended by the agent. Here, the agent controls its movement through 8 available actions (i.e., the King's moves): North (N), South (S), East (E), West (W), North-West (NW), South-West (SW), South-East (SE), and North-East (NE). Every episode terminates either at the allowed time horizon, or when the agent reaches the goal state.

2.1 Grid-world complexity

To test the performance of our active inference agent in a complex stochastic environment, we introduced different complexity levels to the windy grid-world setting (Table 1).

Table 1: Five complexity levels for the windy grid-world task

Level	Wind	Observability	Transition Dynamics
1	Deterministic	Full (MDP)	Known
2	Stochastic	Full (MDP)	Known
3	Deterministic	Full (MDP)	Learned
4	Stochastic	Full (MDP)	Learned
5	Stochastic	Partial (POMDP)	Known

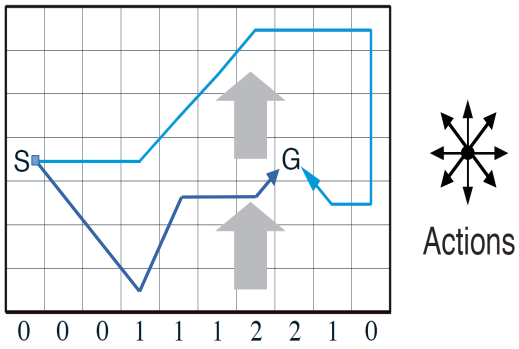


Fig. 1: Windy grid-world task. Here, S and G denote starting and goal locations. On the x-axis, the wind amplitude is shown. This is quantified as the number of unintended additional columns the agent moves during each action e.g., any action in column four results in one unintended shift upwards. There are 8 actions: $N, S, E, W, NW, SW, SE, NE$. We plot sample paths from the start to the goal state in light and dark blue. Notice, the indirect journey to the goal is a consequence of the wind.

Wind properties In a deterministic setting, the amplitude of the wind remains constant. Conversely, in stochastic setting, for windy columns the effect varies by one from the mean values. We consider two settings: medium and high stochasticity. For medium stochasticity, the mean value is observed 70% of the time and similarly 40% of the time in the high stochastic case (Table 2). The adjacent wind values are observed with remaining probabilities. Here, stochasticity is not externally introduced to the system, but it is inbuilt in the transition dynamics $\mathcal{B}$ (Section 3) of the environment.

Table 2: Stochastic nature of wind

Level	Wind amplitude static	Wind amplitude $\pm$ 1
Medium	70% of the time	15% each for ± 1
High	40% of the time	30% each for ± 1

Observability In the fully observable setting, the agent is aware of the current state i.e., there is no ambiguity about the states of affair. We formalise this as a Markov decision processes (MDP). Whereas in the partially observable environment, the agent measures an indirect function of the associated state i.e., current observation. This is used to infer the current state of the agent. We formalise this as a partially observable MDP (POMDP). Specific details of outcome modalities used in the task are discussed in Appendix B.

Transition dynamics known to agent In the known set-up, the agent is equipped with the transition probabilities beforehand. However, if these are not known, the agent begins the trials with a uninformative (uniform) priors and updates its beliefs (Eq.9) using random transitions. Briefly, random actions are sampled and transition dynamics updated to reflect the best explanation for the observations at hand. Here, the learned dynamics are used for planning.

3 Active inference on finite temporal horizons

3.1 Generative model

The generative model is formally defined as a tuple of finite sets (S, O, T, U, B, C, A):

- $s \in S$: states where $S = \{1, 2, 3, ..., 70\}$ and s_1 is a predefined (fixed) start state.
- $o \in O$: where $o = s$, in the fully observable setting, and in partial observability $o = f(s)$[1] .
- $T \in \mathbf{N}^+$, and is a finite time horizon available per episode.
- $a \in U$: actions, where $U = \{N, S, E, W, NW, SW, SE, NE\}$.
- $\mathcal{B}$: encodes the transition dynamics, $P(s_t|s_{t-1}, a_{t-1}, \mathcal{B})$ i.e., the probability that action a_{t-1} taken at state s_{t-1} at time $t-1$ results in s_t at time t.
- $\mathcal{C}$: prior preferences over outcomes, $P(o|\mathcal{C})$. Here, $\mathcal{C}$ preference for the predefined goal-state.
- $\mathcal{A}$: encodes the likelihood distribution, $P(o_\tau|s_\tau, \mathcal{A})$ for the partially observable setting.

Accordingly, the agents generative model is defined as the following probability distribution:

$$P(o_{1:T}, s_{1:T}, a_{1:T-1}, \mathcal{A}, \mathcal{B}, \mathcal{C}) = \tag{1}$$

$$P(\mathcal{A})P(\mathcal{B})P(\mathcal{C})P(s_1)\prod_{\tau=2}^{T} P(s_\tau|s_{\tau-1}, a_{\tau-1}, \mathcal{B})\prod_{\tau=1}^{T} P(o_\tau|s_\tau, \mathcal{A}) \tag{2}$$

3.2 Full observability

Perception: During full observability, states can be directly accessed by agent with known or learned transition dynamics. Then the posterior estimates, $Q(s_{\tau+1}|a_\tau, s_\tau)$, can be directly calculated from $\mathcal{B}$ [11].

$$Q(s_{\tau+1}|a_\tau, s_\tau) = P(s_{\tau+1}|a_\tau, s_\tau, \mathcal{B}). \tag{3}$$

[1] Here, outcomes introduce ambiguity for the agent as similar outcomes map to different (hidden) states. See Appendix B, Table B.1 for implementation details.

Planning: In active inference, expected free-energy ($\mathcal{G}$) [9] is used for planning. For finite temporal horizons, the agent acts to minimise $\mathcal{G}$ [11]. Here, to calculate $\mathcal{G}$ we using the recursive formulation introduced in [10]. This is defined recursively as the immediate expected free energy plus the expected free energy for future actions:

$$\mathcal{G}(a_\tau|s_\tau) = \mathcal{G}(a_{T-1}|s_{T-1}) = D_{KL}[Q(s_T|a_{T-1}, s_{T-1})||C(s_T)] \tag{4}$$

for $\tau = T-1$ and,

$$\mathcal{G}(a_\tau|s_\tau) = D_{KL}[Q(s_{\tau+1}|a_\tau, s_\tau)||C(s_{\tau+1})] \; + \; E_Q\Big[\mathcal{G}(\text{nextstep})\Big] \tag{5}$$

for $\tau = 1, ..., T-2$ [1].

In Eq.5, the second term is calculated as,

$$E_Q\Big[\mathcal{G}(\text{nextstep})\Big] = E_{Q(a_{\tau+1}, s_{\tau+1}|s_\tau, a\tau)}[\mathcal{G}(a_{\tau+1}|s_{\tau+1})]. \tag{6}$$

Prior preference over states are encoded such that the agent prefers to observe itself in the goal state at every time-step. $C(o = \text{goal}) = 1$, and 0 otherwise [2]. In the vector form, the ith element of C, corresponds to ith state in S.

Action selection: A distribution for action selection $Q(a_\tau|s_\tau) > 0$ is defined using expected free energy such that,

$$Q(a_\tau|s_\tau) = \sigma\left(-\mathcal{G}\left(U|s_\tau\right)\right). \tag{7}$$

Here, σ is the softmax function ensuring that components sum to one. At each time-step, actions are samples from:

$$a_t \sim Q(a_t|s_t). \tag{8}$$

Learning transition dynamics: We learn the transition dynamics, $\mathcal{B}$, across time using conjugacy update rules [14,12,9]:

$$b_a = b_a + \sum_{\tau=2}^{t} \sum_{a \epsilon U} \delta_{a,a_\tau} Q(a)\left(s_{a,\tau} \otimes s_{a,\tau-1}\right). \tag{9}$$

Here, $b_a \sim Dir(b; \alpha)$ is the learned transition dynamics updated over time, $Q(a)$ is the probability of taking action a, $s_{a,\tau}$ is the state at time τ as a consequence of action a, $s_{a,\tau-1}$ is the state-vector at time $\tau - 1$ taking action a, and $\otimes$ is the Kronecker-product of the corresponding state-vectors. Furthermore, we also assessed the model accuracy obtained after a given number of trials to update $\mathcal{B}$, when random actions were employed to explore transition dynamics. These learned transitions were used for control in Level-3 and Level-4 of the problem.

[1] First term in Eq.5 does not contribute to solving the problem addressed in the paper. Here, C only accommodates preference to goal-state. However, for a more informed C i.e with preferences for immediate reward maximisation, the term will influence action selection.

[2] The elements in C should be given a finite negligible value while implementation, to avoid divergence of D_{KL} terms in Eq.4 and Eq.5

3.3 Partial observability

We formalise partial observability as a partially observed MDP (POMDP). Here, the agents have access to indirect observations about the environment. Specific details of outcome modalities used in this work are discussed in Appendix B. These outcome modalities are same for many states for e.g., the states 2 and 11 have the same outcome modalities (see Appendix B, Table B.1). Here, we evaluate the ability of active inference agent to perform optimal inference and planning in the face of ambiguity. The critical advancement with sophisticated inference [10] compared to the classical formulation [9] allows us to perform deep-tree search for actions in the future. The agent infers the hidden-states by minimising a functional of its predictive distribution (generative model) of the environment called the variational free-energy. This predictive distribution can be defined as,

$$Q(\vec{s}|\vec{a},\tilde{o}) := \prod_{\tau=1}^{T} Q(s_\tau|a_{\tau-1}, s_{\tau-1}, \tilde{o}). \tag{10}$$

To infer hidden-states from partial observations, the agent engages in minimising variational free energy ($\mathcal{F}$) functional of Q using variational (Bayesian) inference. For a rigorous treatment of it, please refer to [10,11]. In this scheme, actions are considered as random variables at each time-step, assuming successive actions are conditionally independent. This comes with a cost of having to consider many action sequences in time. The search for policies in time is optimised both by restricting the search over future outcomes which has a non-trivial posterior probability (Eg: $> 1/16$) as well as only evaluating policies with significant prior probabilities (Eg: $> 1/16$) calculated from the expected free energy (i.e., Occam's window). In the partially observable setting, the expected free energy accommodates ambiguity in future observations prioritising both preference seeking as well as ambiguity reduction in observations [10].

4 Results

We compare the performance of our active inference agent with a popular reinforcement learning algorithm, Q-learning [13], in Level 1. Q-Learning is a model-free RL algorithm that operates by learning the 'value' of actions at a particular state. It is well suited for problems with stochastic transitions and reward dynamics due to its model-free parametrization. Q-Learning agents are extensively used in similar problem settings and exhibit state-of-the-art (SOTA) performances [13]. To train the Q-learning agents, we used an exploration rate of 0.1, learning rate of 0.5 and discount factor of 1. Training was conducted using 10 different random seeds to ensure unbiased results. The training depth for Q-Learning agents were increased with complexity of the environment.

We instantiate two Q-learning agents, one trained for 500 time-steps (QLearning500) and another for 5000 time-steps (QLearning5K) in Level-1. Both the active inference agent and the QLearning5K agent demonstrate optimal success rate for the time-horizon $T = 8+$ (see Appendix A, Fig.A.1).

Using these baselines from the deterministic environment with known transition dynamics, we compared the performance of the agent in a complex setting with medium and highly stochastic wind (Level 2; Table. 2).Here, the active inference agent is clearly

superior against the Q-Learning agents (Fig. 2 top row). Moreover, they demonstrate better success rates for shorter time-horizons, and 'optimal' action selection. Note, success rate is the percentage of trials for which the agent successfully reached the goal within the allowed time-horizon.

Next, we considered how learning the transition dynamics impacted agent behaviour (Level 3 and 4). Here, we used Eq. 9 for learning the transition dynamics, $\mathcal{B}$. First, the algorithm learnt the dynamics by taking random actions over X steps (for example, X is 5000 time steps in 'SophAgent (5K B-updates)', see Fig. 2 middle row). These learned transition dynamics $\mathcal{B}$ were used (see Fig. 3) by the active inference agent to estimate the action distribution in Eq. 8. Results for level 3 are presented in Appendix A, Fig. A.2. Here, the Q-Learning algorithm with $5,000$ learning steps shows superior performance to the active inference agents. However with longer time horizons, the active inference agent shows competitive performance. Importantly, the active inference agent used self-learned, and imprecise transition dynamics $\mathcal{B}$ in these levels. Level 4 results for medium and highly stochastic setting are presented in Fig. 2 (middle row). For medium stochasticity, the QLearning10K exhibited satisfactory performance, however it failed with zero success rate in the highly stochastic case. This shows the need for extensive training for algorithms like Q-Learning in highly stochastic environments. However, the active inference agent demonstrated at-par performance. Remarkably, the performance was achieved using imprecise (compared to true-model), self-learned transition dynamics ($\mathcal{B}$) (see Fig. 3).

The active inference agent shows superior performance in the highly stochastic environment even with partial observability (Fig. 2, last row). Conversely, excessive training was required for the Q-Learning agent to achieve a high success rate in a medium stochastic environment, but even this training depth led to a zero success rate with high stochasticity. These results present active inference, with a recursively calculated free-energy, as a promising algorithm for stochastic control.

5 Discussion

We explored the utility of the active inference with planning in finite temporal-horizons for five complexity levels of the windy grid-world task. Active inference agents performed at-par, or superior, when compared with well-trained Q-Learning agents. Importantly, in the highly stochastic environments the active inference agent showed clear superiority over the Q-Learning agents. The higher success rates at lower time horizons demonstrated the 'optimality' of actions in stochastic environments presented to the agent. Additionally, this performance is obtained with no specifications of acceptable policies. The total number of acceptable policies scale exponentially with the number of available actions and time-horizon. Moreover, the Level 4 & 5 results demonstrate the need for extensive training for the Q-Learning agents when operating in stochastic environments. We also demonstrated the ability of the active inference agents to achieve high success rate even with self-learned, but sub-optimal, transition dynamics. Methods to equip the agent to learn both transition-dynamics $\mathcal{B}$ and outcome-dynamics $\mathcal{A}$ for a partially observable setting have been previously explored [14,9]. For a stochastic setting, we leave their implementation for future work.

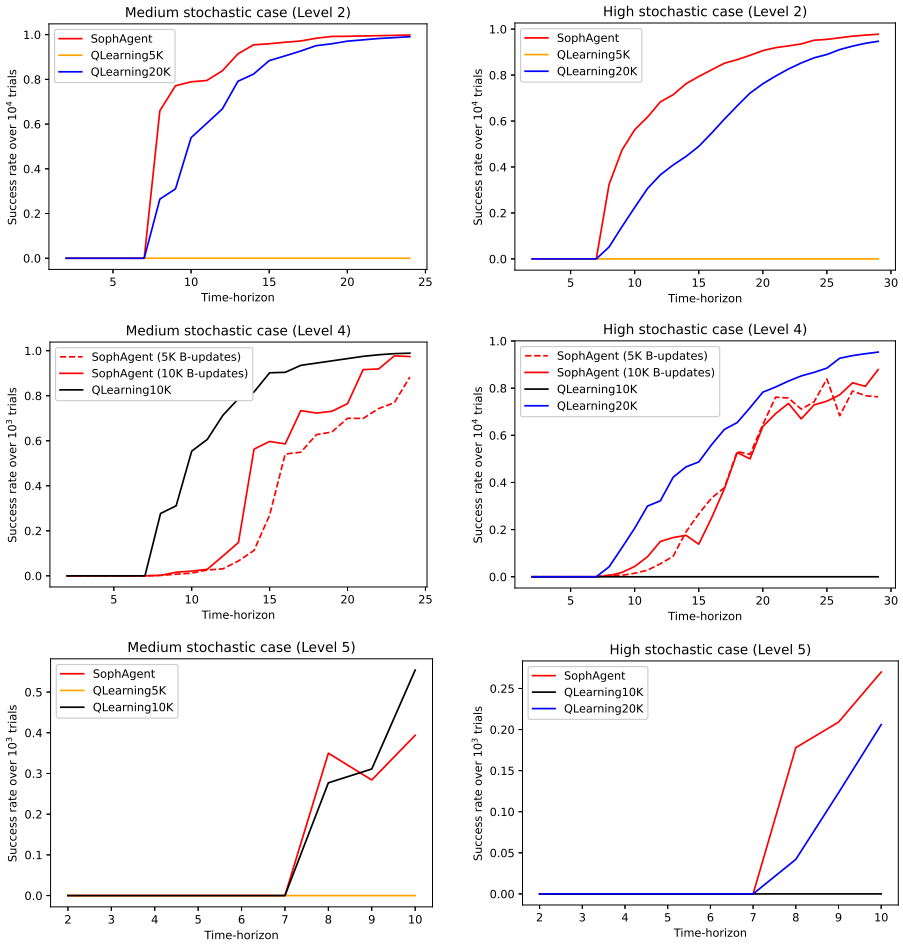


Fig. 2: Stochastic environments: Performance comparison of agents in Level-2 (top row), Level-4 (middle row), and Level-5 (last row) of windy grid-world task for medium-stochastic (left column) and high-stochastic (right column) environments, respectively. Here, x-axis denotes time horizon and y-axis the success rate over multiple-trials. 'SophAgent' represents the active inference agent, 'QLearning5K' represents Q-learning agent trained for $5,000$ time-steps, 'QLearning10K' for the 'Q-learning agent trained for $10,000$ time-steps, and 'QLearning20K' for the Q-learning agent trained for $20,000$ time-steps. Each agent was trained using 10 different random seeds. 'SophAgent (5K B-updates)' and SophAgent (10K B-updates) refers to active inference agent using self-learned transition dynamics $\mathcal{B}$ with 5000 and 10000 updates respectively.

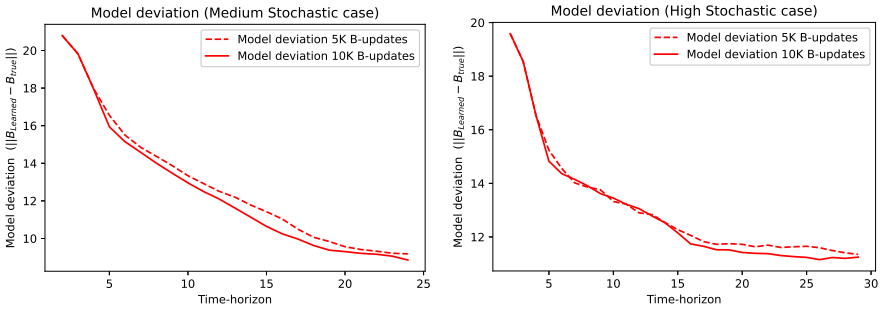


Fig. 3: Accuracy of learned dynamics in terms of deviation from true transition dynamics in Level-4 A: Medium stochastic case B: High stochastic case

The limitation yet to be addressed is the time consumed for trials in active inference. Large run-time restricted analysis for longer time horizons in Level 5. Deep learning approaches using tree searches, for representing policies were proposed recently [15,16,17], may be useful in this setting. We leave run-time analysis and optimisation for more ambitious environments for future work. Also, comparing active inference to model based RL algorithms like Dyna-Q [13] and control as inference approaches [18] is a promising direction to pursue.
We conclude that the above results place active inference as a promising algorithm for stochastic-control.

Software note The environments and agents were custom written in Python for fully observable settings. The script 'SPM_MDP_VB_XX.m' available in SPM12 package was used in the partially observable setting. All scripts are available in the following link: `https://github.com/aswinpaul/iwai2021_aisc`.

Acknowledgments AP acknowledges research sponsorship from IITB-Monash Research Academy, Mumbai and Department of Biotechnology, Government of India. AR is funded by the Australian Research Council (Refs: DE170100128 & DP200100757) and Australian National Health and Medical Research Council Investigator Grant (Ref: 1194910). AR is a CIFAR Azrieli Global Scholar in the Brain, Mind & Consciousness Program. AR and NS are affiliated with The Wellcome Centre for Human Neuroimaging supported by core funding from Wellcome [203147/Z/16/Z].

References

1. Friston, K.: The free-energy principle: a unified brain theory?. Nat Rev Neuroscience 11, 127–138 (2010).
2. Kaplan, Raphael and Friston, Karl J: Planning and navigation as active inference. Biological cybernetics **112**(4), 323–343 (2018).
3. Kuchling, Franz and Friston, Karl and Georgiev, Georgi and Levin: Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. Physics of life reviews, (2019)
4. Oliver, Guillermo and Lanillos, Pablo and Cheng, Gordon: Active inference body perception and action for humanoid robots. arXiv preprint arXiv:1906.03022, (2019)

5. Rubin, Sergio and Parr, Thomas and Da Costa, Lancelot and Friston, Karl: Future climates: Markov blankets and active inference in the biosphere. Journal of the Royal Society Interface **17**(172), (2020)
6. Deane, George and Miller, Mark and Wilkinson, Sam: Losing Ourselves: Active Inference, Depersonalization, and Meditation. Frontiers in Psychology, (2020)
7. Friston KJ, Daunizeau J, Kiebel SJ.: Reinforcement Learning or Active Inference? PLoS ONE 4(7): e6421, (2009) https://doi.org/10.1371/journal.pone.0006421
8. Friston, Karl and Samothrakis, Spyridon and Montague, Read: Active inference and agency: optimal control without cost functions. Biological cybernetics **106**(8), 523-541 (2012)
9. Noor Sajid, Philip J. Ball, Thomas Parr, Karl J. Friston.: Active Inference: Demystified and Compared. Neural Computation 33 (3), 674–712 (2021)
10. Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, Thomas Parr: Sophisticated Inference. Neural Comput 2021; 33 (3), 713–763 (2021).
11. Lancelot Da Costa and Noor Sajid and Thomas Parr and Karl Friston and Ryan Smith:, The relationship between dynamic programming and active inference: the discrete, finite-horizon case.:, arXiv.2009.08111, (2020).
12. Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K.: Active inference on discrete state-spaces: a synthesis", arXiv e-prints, (2020).
13. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press (2018).
14. Friston, Karl and FitzGerald, Thomas and Rigoli, Francesco and Schwartenbeck, Philipp and Pezzulo, Giovanni: Active inference: a process theory. Neural computation **29**(1), 1–49 (2017)
15. Fountas, Zafeirios and Sajid, Noor and Mediano, Pedro AM and Friston, Karl: Deep active inference agents using Monte-Carlo methods. arXiv preprint arXiv:2006.04176, (2020)
16. Çatal, Ozan and Nauta, Johannes and Verbelen, Tim and Simoens, Pieter and Dhoedt, Bart: Bayesian policy selection using active inference. arXiv preprint arXiv:1904.08149, (2019)
17. van der Himst, Otto Lanillos, P.: Deep Active Inference for Partially Observable MDPs. In: Verbelen, Tim and Lanillos, Pablo and Buckley, Christopher L. and De Boom, Cedric (eds.), Active Inference, pp. 61–71, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-64919-7
18. Millidge, Berenand Tschantz, Alexanderand Seth, Anil K. and Buckley, Christopher L.: On the Relationship Between Active Inference and Control as Inference. In: Verbelen, Tim and Lanillos, Pablo and Buckley, Christopher L. and De Boom, Cedric (eds.), Active Inference, pp. 3–11, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-64919-7

Supplementary information

A Results Level-1 and Level-3 (Non-stochastic settings)

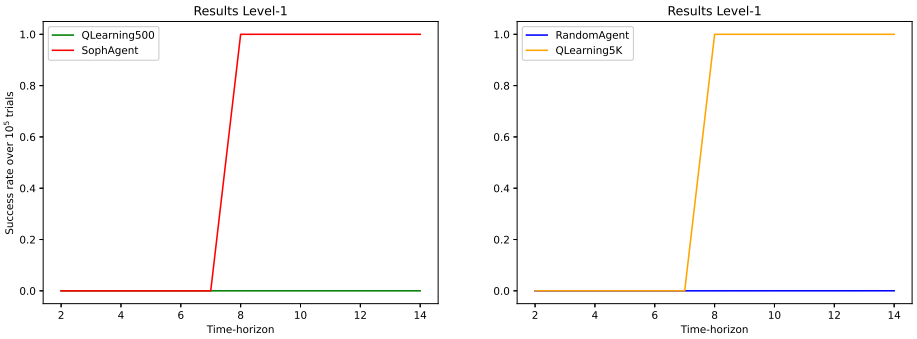


Fig. A.1: Performance comparison of agents in Level-1 of windy grid-world task. 'RandomAgent' refers to a naive-agent that takes all actions with equal probability at every time step.

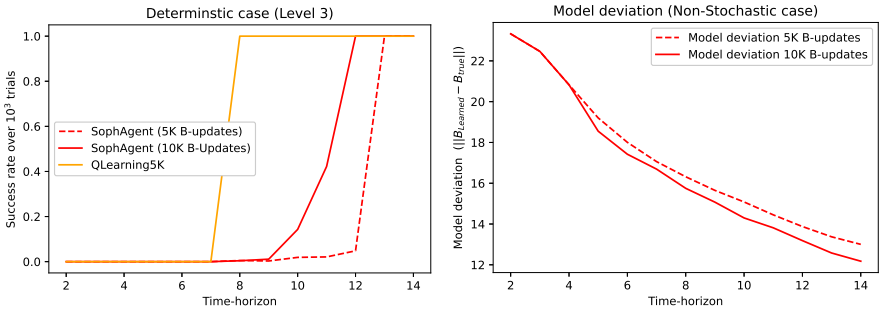


Fig. A.2: A: Performance comparison of active inference agents with learned B using 5000 and 10000 updates respectively to Q-Learning agent in Level-3. 'Q-Learning5K' stands for Q-Learning agent trained for 5000 time steps using 10 different random seeds. B: Accuracy of learned dynamics in terms of deviation from true dynamics.

B Outcome modalities for POMDPs

In the partially observable setting, we considered two outcome modalities and both of them were the function of 'side' and 'down' coordinates defined for every state in Fig. 1. Examples of the coordinates and modalities are given below. First outcome modality is the sum of co-ordinates and second modality is the product of coordinates.

Table B.1: Outcome modalities specifications

State	Down coordinate (C1)	Side coordinate (C2)	Outcome-1 (C1+C2)	Outcome-2 (C1*C2)
1	1	1	2	1
2	1	2	3	2
.	.	.	.	.
11	2	1	3	2
.	.	.	.	.
31	4	1	5	4
38	4	8	12	32
.	.	.	.	.

These outcome modalities are similar for many states (for e.g., states 2 and 11 have the same outcome modalities (see Tab. B.1)). The results demonstrates the ability of active inference agent to perform optimal inference and planning in the face of ambiguity. One of the output from 'SPM_MDP_VB_XX.m' is 'MDP.P'. 'MDP.P' returns the action probabilities an agent will use for a given POMDP as input at each time-step. This distribution was used to conduct multiple trails to evaluate success rate of the active inference agent.

Towards Stochastic Fault-tolerant Control using Precision Learning and Active Inference

Mohamed Baioumy[1], Corrado Pezzato[2], Carlos Hernández Corbato[2], Nick Hawes[1], and Riccardo Ferrari[3]

[1] Oxford Robotics Institute, University of Oxford
{mohamed, nickh}@robots.ox.ac.uk
[2] Cognitive Robotics, Delft University of Technology
[3] Delft Center for Systems and Control, Delft University of Technology
{c.pezzato, c.h.corbato, r.ferrari, m.wisse}@tudelft.nl

Abstract. This work presents a fault-tolerant control scheme for sensory faults in robotic manipulators based on active inference. In the majority of existing schemes a binary decision of whether a sensor is healthy (functional) or faulty is made based on measured data. The decision boundary is called a threshold and it is usually deterministic. Following a faulty decision, fault recovery is obtained by excluding the malfunctioning sensor. We propose a stochastic fault-tolerant scheme based on active inference and precision learning which does not require a priori threshold definitions to trigger fault recovery. Instead, the sensor precision, which represents its health status, is learned online in a model-free way allowing the system to gradually, and not abruptly exclude a failing unit. Experiments on a robotic manipulator show promising results and directions for future work are discussed.

1 Introduction

Safety is paramount for autonomous systems designed for operating in the real world. External dangers in the environment such as steep and slippery terrain encountered by planetary rovers [14] can compromise entire missions. In addition to external dangers, internal system components can also fail and possibly lead to dangerous outcomes if a proper fault-tolerant (FT) control scheme is not present. Building systems that are robust to the presence of faulty components, such as sensors and actuators, is addressed in the FT literature [23,31,8]. Generally speaking, FT control consists of *fault detection*, which provides a signal representing whether a system component is faulty; *fault isolation*, which identifies the exact faulty component, and *fault recovery*, which typically contains a switching or a re-tuning procedure of the running controllers to accommodate for the fault.
Several methods are available for fault detection, but model-based methods are among the most powerful and appealing, as they provide theoretical guarantees [8]. These methods rely on monitoring system outputs using mathematical models to generate 'symptoms' called *residual signals*. These signals are then compared to carefully designed detection

thresholds: the sensor is 'faulty' if a threshold is exceeded or 'healthy' otherwise. To recover from a fault, the recovery actions are usually performed through controller reconfiguration, that entails adapting the controller parameters, or switching to another controller or to backup sensors and actuators [21]. When modelling external dangers or monitoring faulty systems, robust detection thresholds are essential. Robust thresholds used in existing work (such as [7] or [31]) are often *deterministic*, but this is sub-optimal. For instance, if the safety threshold for a rover on a slippery terrain slope is 15 degrees, this means that a slope of 14.9 is safe but 15.1 is unsafe. Additionally, a slope of 15.1 degrees and 40 degrees are 'equally unsafe'.

In this paper we build upon two ideas in the literature. First, the usage of a stochastic fault tolerant formulation (e.g. [9,29]). This allows the agent to overcome the issues mentioned above. Additionally, we leverage an unbiased active inference controller (u-AIC) [3], evolved from previous active inference controllers (AIC) [6,1,24]. Active inference is a promising framework for FT control which has already been shown to facilitate fault-detection, isolation and recovery for robotic systems with sensory faults [3,25].

Besides fault tolerance, active inference showed promising performance in many control and state-estimation problems in robotics [15,16]. Particularly interesting are the works on robot arm control [24,22,28], which highlighted the adaptive properties of active inference. Active inference also shares similarities with the control as inference framework [17]. A more extensive analysis of active inference and its relation to control as inference can be found in [19,13].

The main contribution of this paper is a FT controller for robot manipulators with sensory faults based on unbiased active inference with a stochastic decision boundary. Unlike previous work [3], here we model the precision (inverse covariance) of each sensor in our system and determine the probability of the sensor being healthy to be proportional to its precision. Our approach allows for fault-tolerant behaviour without needing any threshold definition a priori, and without the need to design additional ad-hoc recovery mechanisms. Finally, this work can be used stand-alone or in conjunction with other methods for fault-detection and isolation in order to estimate the faults.

2 Problem statement and background

The FT scheme in this paper is derived for a class of systems, namely serial robot manipulators equipped with sensors for joint position and velocity, and end-effector location. In the following, the problem and the setup are described, and some background knowledge on u-AIC for torque control from [3] is presented.

Problem Setup. Consider a robotic manipulator with state $\boldsymbol{x}$ comprising of its joint positions and velocities $\boldsymbol{x} = [q\ \dot{q}]^\top$. The available sensors provide noisy joint position and velocities $\boldsymbol{y}_q$, $\boldsymbol{y}_{\dot{q}}$ readings. In addition, the end-effector Cartesian position $\boldsymbol{y}_v$ is available through a visual

sensor. The system's output is represented by $\boldsymbol{y} = [\boldsymbol{y}_q,\ \boldsymbol{y}_{\dot{q}},\ \boldsymbol{y}_v] \in \mathbb{R}^d$. The proprioceptive sensors and the visual sensor are affected by zero mean Gaussian noise $\boldsymbol{\eta} = [\boldsymbol{\eta}_q, \boldsymbol{\eta}_{\dot{q}}, \boldsymbol{\eta}_v]$. Additionally, the visual sensor is affected by barrel distortion. The system is controlled through an u-AIC [3] which steers the robot arm to a (changing) desired configuration in joint space $\boldsymbol{\mu}_d$, providing the control input $\boldsymbol{u} \in \mathbb{R}^m$ as torques to the joints.

Background: Unbiased Active Inference controller In this section we briefly describe the u-AIC as introduced in [3], to which an interested reader is referred for more details on the derivations of the following equations. The novel FT method presented in this paper in Sec. 3 builds upon the u-AIC, but instead of employing an ad-hoc hard update of the precision of a faulty sensor after fault detection, it relies on online precision learning during operations.
Let us consider $\boldsymbol{x} = [q\ \dot{q}]^\top$ and let us define a probabilistic model where actions are modelled explicitly:

$$p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}_v, \boldsymbol{y}_q, \boldsymbol{y}_{\dot{q}}) = \underbrace{p(\boldsymbol{u}|\boldsymbol{x})}_{control}\ \underbrace{p(\boldsymbol{y}_v|\boldsymbol{x})p(\boldsymbol{y}_q|\boldsymbol{x})p(\boldsymbol{y}_{\dot{q}}|\boldsymbol{x})}_{observation\ model}\ \underbrace{p(\boldsymbol{x})}_{prior} \tag{1}$$

Note that with the u-AIC the information about the desired goal to be reached is encoded in the distribution $p(\boldsymbol{u}|\boldsymbol{x})$. In this paper, as in [1], we assume that an accurate dynamic model of the system is not available to keep the solution system agnostic and to highlight once again the adaptability of the controller.
The u-AIC aims at finding the posterior over states as well as the posterior over actions $p(\boldsymbol{x}, \boldsymbol{u}|\boldsymbol{y}_v, \boldsymbol{y}_q)$. The posteriors are approximated using a variational distribution $Q(\boldsymbol{x}, \boldsymbol{u})$. We can make use of the mean-field assumption ($Q(\boldsymbol{x}, \boldsymbol{u}) = Q(\boldsymbol{x})Q(\boldsymbol{u})$) and the Laplace approximation, and assume the posterior over the state $\boldsymbol{x}$ Gaussian with mean $\boldsymbol{\mu}_x$ [12]. Similarly for the actions, the posterior $\boldsymbol{u}$ is assumed Gaussian with mean $\boldsymbol{\mu}_u$. By defining the Kullback-Leibler divergence between the variational distribution and the true posterior, one can derive an expression for the so-called free-energy F as [3]:

$$F = -\ln p(\boldsymbol{\mu}_u, \boldsymbol{\mu}_x, \boldsymbol{y}_v, \boldsymbol{y}_q, \boldsymbol{y}_{\dot{q}}) + C \tag{2}$$

Considering eq. (1) and assuming Gaussian distributions, F becomes:

$$\begin{aligned} F = \frac{1}{2}(&\boldsymbol{\varepsilon}_{y_q}^\top \Sigma_{y_q}^{-1} \boldsymbol{\varepsilon}_{y_q} + \boldsymbol{\varepsilon}_{y_{\dot{q}}}^\top \Sigma_{y_{\dot{q}}}^{-1} \boldsymbol{\varepsilon}_{y_{\dot{q}}} + \boldsymbol{\varepsilon}_{y_v}^\top \Sigma_{y_v}^{-1} \boldsymbol{\varepsilon}_{y_v} \\ &+ \boldsymbol{\varepsilon}_x^\top \Sigma_x^{-1} \boldsymbol{\varepsilon}_x + \boldsymbol{\varepsilon}_u^\top \Sigma_u^{-1} \boldsymbol{\varepsilon}_u + \ln|\Sigma_u \Sigma_{y_q} \Sigma_{y_{\dot{q}}} \Sigma_{y_v} \Sigma_x|) + C, \end{aligned} \tag{3}$$

The terms $\boldsymbol{\varepsilon}_{y_q} = \boldsymbol{y}_q - \boldsymbol{\mu}$, $\boldsymbol{\varepsilon}_{y_{\dot{q}}} = \boldsymbol{y}_{\dot{q}} - \boldsymbol{\mu}'$, $\boldsymbol{\varepsilon}_{y_v} = \boldsymbol{y}_v - \boldsymbol{g_v}(\boldsymbol{\mu})$ are the sensory prediction errors respectively for position, velocity, and visual sensory inputs. The controller represents the states internally as $\boldsymbol{\mu}_x = [\boldsymbol{\mu},\ \boldsymbol{\mu}']^\top$. The relation between internal state and observation is expressed through the generative model of the sensory input $\boldsymbol{g} = [\boldsymbol{g}_q,\ \boldsymbol{g}_{\dot{q}},\ \boldsymbol{g}_v]$. Position and velocity encoders directly measure the state, thus $\boldsymbol{g}_q$ and $\boldsymbol{g}_{\dot{q}}$ are linear (identity) mappings. To define $\boldsymbol{g}_v$, instead, we use a *Gaussian Process*

Regression (GPR). This is particularly useful because we can model the noisy and distorted sensory input from the camera, and at the same time we can compute a closed form for the derivative of the process with respect to the beliefs $\boldsymbol{\mu}$, required for the state update laws. For details, see [3].

Additionally, $\boldsymbol{\varepsilon}_u$ is the prediction error on the control action while $\boldsymbol{\varepsilon}_x$ is the prediction error on the state. The latter is computed considering a prediction of the state $\hat{\boldsymbol{x}}$ at the current time-step such that $\boldsymbol{\varepsilon}_x = (\boldsymbol{\mu}_x - \hat{\boldsymbol{x}})$. The prediction is a deterministic value $\hat{\boldsymbol{x}} = [\hat{q}\ \hat{\dot{q}}]^\top$ which can be computed in the same fashion as the prediction step of, for instance, a Kalman filter. The prediction is approximated propagating forward in time the current state belief using the following simplified discrete time model:

$$\hat{x}_{k+1} = \begin{bmatrix} I & I\Delta t \\ 0 & I \end{bmatrix} \mu_{x,k} \tag{4}$$

where I represents an unitary matrix of suitable size. This form assumes that the position of each joint is thus computed as the discrete time integral of the velocity, using a first-order Euler scheme. This approximation can be avoided if a better dynamic model of the system is available, and in that case predictions can be made using the model itself. Finally, by choosing the distribution $p(\boldsymbol{u}|\boldsymbol{x})$ to be Gaussian with mean $f^*(\boldsymbol{\mu}_x, \boldsymbol{\mu}_d)$, we can steer the systems toward the target $\boldsymbol{\mu}_d$ without biasing the state estimation. This results in $\boldsymbol{\varepsilon}_u = (\boldsymbol{\mu}_u - f^*(\boldsymbol{\mu}_x, \boldsymbol{\mu}_d))$.

In the u-AIC state-estimation and control are achieved using gradient descent the free-energy. This leads to:

$$\dot{\boldsymbol{\mu}}_u = -\kappa_u \frac{\partial F}{\partial \boldsymbol{\mu}_{\boldsymbol{u}}}, \quad \dot{\boldsymbol{\mu}}_x = -\kappa_\mu \frac{\partial F}{\partial \boldsymbol{\mu}_{\boldsymbol{x}}}, \tag{5}$$

where κ_u and κ_μ are the gradient descent step sizes.

3 Precision Learning for fault-tolerant control

In previous work [3], the u-AIC is used in combination with an established FT approach to achieve fault detection and recovery. In particular, the sensory prediction errors in the free-energy are used as residual signals for fault detection purposes. The statistical properties of the residuals are analysed offline and healthy boundaries are defined. At runtime, a healthy residual set is computed and if the current residual is outside the admissible set, the relative sensor is marked as faulty. When a fault is detected, the precision (or inverse covariance) of the sensor is abruptly set to zero, that is $P = \Sigma^{-1} = 0$, to exclude that sensor from the optimization of the free-energy. This idea is summarised in Fig. 1.

In this work, we propose a different approach to achieve fault recovery through online precision learning with u-AIC instead ad-hoc hard switches in the controller's parameters. Fig. 2 shows thee difference with respect to [3].

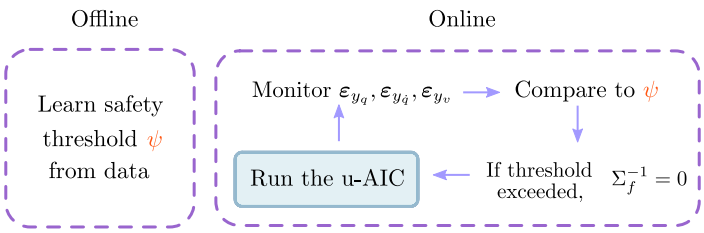


Fig. 1. Fault-tolerant pipeline from [3]. The term Σ_f^{-1} represents the precision of the detected faulty sensor.

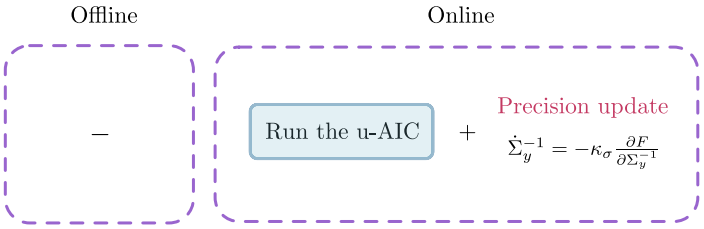


Fig. 2. New fault-tolerant pipeline with precision learning, in contrast to previous work [3] from Fig. 1.

Learning sensory precision. For a sensor y, we can update an inverse precision matrix Σ_y^{-1} using gradient descent on F as done in [2,1]:

$$\dot{\Sigma}_y^{-1} = -\kappa_\sigma \frac{\partial F}{\partial \Sigma_y^{-1}}. \tag{6}$$

However, we need to ensure that precision remains a positive number. Performing gradient descent does not inherently guarantee that.
First, consider a one-dimensional problem where state x and observation y are scalars. The observations is affected by zero-mean Gaussian noise with a variance of σ^2 (also a scalar). The scalar precision is defined as the inverse variance $\omega = 1/\sigma^2$. As explained, performing gradient descent on the free-energy with respect to ω may result in it being negative. A simple solution is to perform a reparameterization with a strictly positive function such as an exponential. I.e. we assume that $\omega = \exp \zeta$ and we perform gradient descent on ζ:

$$\dot{\zeta} = -\kappa_\zeta \frac{\partial F}{\partial \zeta} \tag{7}$$

where κ_ζ is the gradient step-size. Another way is to set a lower bound on the variance (as done in [5]). Both methods ensure the variance being positive.

Diagonal precision matrix.
Guaranteeing a positive semi-definite matrix in an n-dimensional case is not as straightforward. However, in the context of a robotic manipulator, one can reasonably assume that the observation noise on each sensor is

independent [22,24,1]. This means that the covariance (and precision) matrices are diagonal.

$$P = \begin{bmatrix} \omega_1 & & & \\ & \omega_2 & & \\ & & ... & \\ & & & \omega_n \end{bmatrix}$$

Given this assumption, every element on the diagonal is positive and can be updated in the same fashion as the scalar case (Eq. (7)).

Fault-tolerant control as precision learning. Consider the sum of the sensory prediction errors in the free-energy from eq. (3):

$$F = \frac{1}{2}(\boldsymbol{\varepsilon}_{y_q}^{\top} \Sigma_{y_q}^{-1} \boldsymbol{\varepsilon}_{y_q} + \boldsymbol{\varepsilon}_{y_{\dot{q}}}^{\top} \Sigma_{y_{\dot{q}}}^{-1} \boldsymbol{\varepsilon}_{y_{\dot{q}}} + \boldsymbol{\varepsilon}_{y_v}^{\top} \Sigma_{y_v}^{-1} \boldsymbol{\varepsilon}_{y_v} + ...) + C, \qquad (8)$$

Intuitively, when a sensor is faulty, the related sensory prediction error will necessarily be higher since sensory readings and internal beliefs will drift away. After a fault, the estimated precision through our precision learning scheme will be much lower than the original $P = \Sigma^{-1}$. Thus its weight in the free-energy F, and so in the state-estimation and control equations as in eq. (5) will naturally become lower than the other healthy sensors. Its weight essentially adjusts *proportionally to the degree of the sensor being faulty*. Note that this allows for automatic fault recovery but it does not provide explicit fault detection. In case the latter is needed for a potential user or an additional supervisory system, traditional techniques can be used as the one presented in [3] in conjunction with precision learning.

FT control for sensory faults can now be done using precision learning in several ways. The first way is to use it as a stand-alone and activate precision learning for all sensors during operation. In this case, no other methods are needed, no thresholds are designed and the recovery emerges naturally. As mentioned before, the drawback is that the users can not be 'alerted' for the presence of a fault (since there is no explicit fault-detection). The second way, which addresses this issue, is to use an established algorithm for fault detection (such as the one presented in [3]) and then, only after a fault is detected, allow precision update.

Interestingly, performing precision learning as presented in this section can make the state-estimation noisier since the agents only relies on the current observation (rather than a batch of last k observations) for the update and both the uncertainty of the state and precision are not quantified. An additional approach would then be to consider the last k observations for the update, but this is out of the scope of this work.

To summarise, the precision learning in this paper can either be activated at all times or *only after a fault is detected*. Activating the precision learning at all times with a small step-size for the gradient seems to work best.

4 Results

We apply the methods in Sec. 3 on a 2-DOF robotic manipulator. We test three scenarios: a) precision learning at all times, b) precision learning only when a fault is detected and c) a deterministic update as done in [3]. Note that the latter has access to a model and uses data to determine a threshold offline. This is not the case for the first two options where only model-free precision learning is performed. The results are summarized in the Table 1. In the simulations, the sensors are injected with zero-mean

	Joints with encoder fault	Joints without encoder fault
No fault-tolerance	0.0036	0.0020
PL at all time	5.422 e-5	4.527 e-5
PL + fault-detection	6.097 e-5	4.134 e-5
Deterministic fault recovery	**0.5946 e-5**	**0.3579 e-5**

Table 1. Mean Squared Error (MSE) for different methods of fault-tolerant control. PL indicates precision learning

Gaussian noise. The standard deviation of the noise for encoders and velocity sensors is set to $\sigma_q = \sigma_{\dot{q}} = 0.001$, while the one for the camera is set to $\sigma_v = 0.01$. The camera is also affected by barrel distortion with coefficients $K_1 = -1.5e^{-3}$, $K_2 = 5e^{-6}$, $K_3 = 0$ (values are similar to work from [18,27]).

The agents starts in configuration $\boldsymbol{x}_0$, then moves to the targets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. At $t = 8s$ a fault is injected. The encoder fault is such that the output related to the first joint freezes. For a discrete step k it holds then $\boldsymbol{y}_q(k) = [q_1(k_f),\ q_2(k)]^\top$ for $k \geq k_f$ and $k_f = 8$. The fault detection and recovery of such a fault, as well as the system's response, are reported below in Fig. 3.

As seen in Fig. 3, the system is not able to reach the set-point after the occurrence of the fault if online precision update is not allowed. The robot arm reaches a different configuration to minimise the free-energy, which is built fusing the sensory information from the (faulty) encoders and the (healthy) camera. However, when the faulty encoder is adjusted using precision learning, the agent is able to reach the final configuration. Fig. 3 reports the results when precision learning is being done during the full operational time. Alternatively, one could only use precision leaning when a fault is detected. This yields a response that is almost identical. The Mean Squared Error (MSE) between the belief and the true position $(\boldsymbol{\mu}_x - \boldsymbol{x})$ is computed on a sample of test runs and reported in the Table 1. The results are reported for both the joint whose encoder is faulty, and joints with healthy encoders. In both cases, hard update of the precision to zero has the lowest MSE; however, the approaches based on precision learning do not require any previous information or a threshold definition thus it is simpler to implement. Yet, precision learning has a satisfactory performance while accommodating a sensory fault.

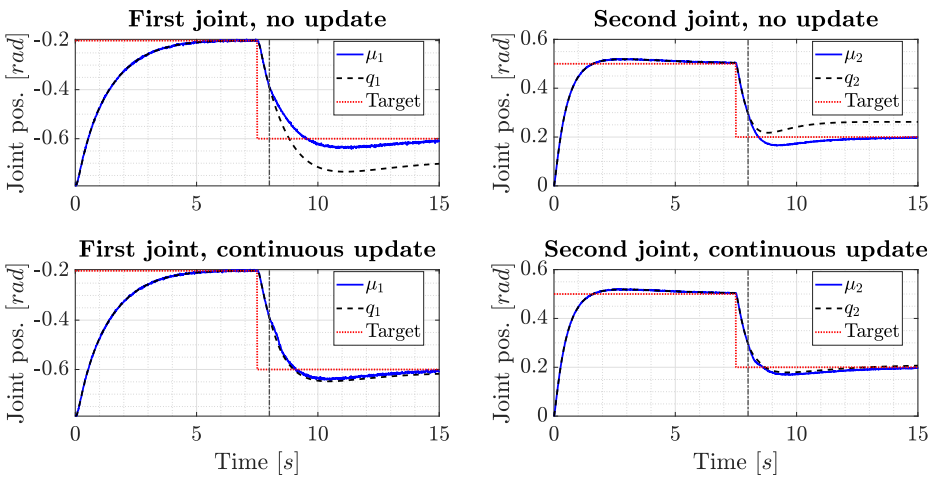


Fig. 3. System's response in the case of encoder fault with and without precision learning applied at all times. The fault is injected at $t = 8s$ and indicated with a dot-dashed line.

5 Improving precision learning: a discussion

In this paper, we perform a simple modification to the unbiased active inference controller: adding precision learning for all sensors. We show that this results in stochastic fault-tolerance to sensory faults, i.e. the precision of a faulty sensor will decrease automatically making its relative weight in the control and estimation laws smaller. This eliminates the need to learn a threshold from data offline. Additionally, no ah-hoc recovery action is required. The controller automatically adjusts to the new precision.

In the experiments, we compared precision learning to a state-of-the-art method. Precision learning was an order of magnitude worse in performance but still satisfactory. Note that precision learning did *not* require any data or training offline to determine thresholds or recovery strategies. Finally, precision learning performs stochastic fault-detection rather than deterministic.

Most importantly, this approach based on precision learning can be improved in many ways. First, rather than computing a point-mass estimate, we can explicitly model the precision as a random variable and perform inference on it.

We can perform Bayesian inference by modelling the precision as a random variable and computing a posterior over it. In the one dimensional case we use a Gamma prior on the precision ω as

$$\Gamma(\omega; a, b) = \frac{b^a}{\Gamma(a)}\omega^{a-1}e^{-\omega b}.$$

Given that the observation model is Gaussian, this choice is beneficial since it is the conjugate prior [4,20], where a and b are the parameters of

the distribution and $\Gamma(a) = (a-1)!$ is a factorial function. For example, $\Gamma(5) = 4! = 24$. Now to compute the posterior, we multiply the prior with the Gaussian likelihood model of $p(y|\omega)$ and obtain the posterior which is also a Gamma distribution as shown below.

$$p(\omega) = \Gamma(\omega; a, b) \propto \omega^{a-1} e^{-\omega b}$$

$$p(\omega|y) \propto p(y|\omega)p(\omega) \propto \omega^{0.5+a-1} e^{-\omega(b+\frac{(y-C)^2}{2})}$$

$$p(\omega|y) = \Gamma(\omega; a + \frac{1}{2}, b + \frac{(y-C)^2}{2})$$

The last equation shows a simple update rule to modify the belief over the precision for every data point. In the optimization for the state, the following quantities are used: expected precision $\mathbb{E}[\omega] = a/b$, $Mode[\omega] = (a-1)/b$ and $Var[\omega] = a/b^2$. In the n-dimensional case, the same procedure can be done but with a Wishart distribution rather than a Gamma. Additionally, we could use a batch of k observation to learn the precision rather than just one observation. Many approaches for covariance/precision estimation have been successful in robotics e.g. [32,26,33,30]. Additionally, many other approaches within the active inference literature can be used for effective precision learning such as dynamic expectation maximization (DEM) [11,10]. These will be explored and compared in future work.

6 Conclusions

This paper presents a fault-tolerant controller based on active inference. We model the precision (inverse covariance) of each sensor in our system and determine the probability of the sensor being healthy to be proportional to its precision. Rather than reasoning about whether a sensor is faulty or not, we reason about the degree to which the sensor is faulty. We present gradient based approaches to approximate the precision matrices of the system. The results show that the precision learning is a promising approach for fault-tolerant control. It allows for robust behaviour without needing any threshold definition a priori, without designing additional ad-hoc recovery mechanisms, and can be used stand-alone or in conjunction with other methods. The results using precision learning was satisfactory but an order of magnitude away from the to state-of-the-art. However, precision learning was not trained on data offline and performs a stochastic update. Bayesian methods can be used to improve the performance of the approach. Additionally, in all cases regarding precision learning, the performance can be improved by considering the last k observations rather than just one. Future work will address this.

References

1. Baioumy, M., Duckworth, P., Lacerda, B., Hawes, N.: Active inference for integrated state-estimation, control, and learning. In: Proc of IEEE Int. conference on robotics and automation (ICRA) (2021)

2. Baioumy, M., Mattamala, M., Duckworth, P., Lacerda, B., Hawes, N.: Adaptive manipulator control using active inference with precision learning. In: UKRAS (2020)
3. Baioumy, M., Pezzato, C., Ferrari, R., Corbato, C.H., Hawes, N.: Fault-tolerant control of robotic systems with sensory faults using unbiased active inference. In: European Control Conference (ECC) (2021)
4. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
5. Bogacz, R.: A tutorial on the free-energy framework for modelling perception and learning. Journal of mathematical psychology **76**, 198–211 (2017)
6. Buckley, C.L., Kim, C.S., McGregor, S., Seth, A.K.: The free energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology **81**, 55–79 (2017)
7. Budd, M., Lacerda, B., Duckworth, P., West, A., Lennox, B., Hawes, N.: Markov decision processes with unknown state feature values for safe exploration using gaussian processes. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020)
8. Chen, J., Patton, R.J.: Robust model-based fault diagnosis for dynamic systems. Springer Science & Business Media, LLC (1999)
9. Fang, S., Blanke, M., Leira, B.J.: Mooring system diagnosis and structural reliability control for position moored vessels. Control Eng. Practice **36**, 12–26 (2015)
10. Friston, K., Stephan, K., Li, B., Daunizeau, J.: Generalised filtering. Mathematical Problems in Engineering **2010** (2010)
11. Friston, K.J., Trujillo-Barreto, N., Daunizeau, J.: Dem: a variational treatment of dynamic systems. Neuroimage **41**(3), 849–885 (2008)
12. Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W.: Variational free energy and the Laplace approximation. Neuroimage **34(1)**, 220–234 (2007)
13. Imohiosen, A., Watson, J., Peters, J.: Active inference or control as inference? a unifying view. In: International Workshop on Active Inference. pp. 12–19. Springer (2020)
14. Inotsume, H., Kubota, T., Wettergreen, D.: Robust path planning for slope traversing under uncertainty in slip prediction. IEEE Robotics and Automation Letters **5**(2), 3390–3397 (2020)
15. Lanillos, P., Cheng, G.: Adaptive robot body learning and estimation through predictive coding. In: IROS (2018)
16. Lanillos, P., G.Cheng: Active inference with function learning for robot body perception. In: Int. Workshop on Continual Unsupervised Sensorimotor Learning (ICDL-Epirob) (2018)
17. Levine, S.: Reinforcement learning and control as probabilistic inference: Tutorial and review. arXiv preprint arXiv:1805.00909 (2018)
18. Marshall, M., Lipkin, H.: Kalman filtering visual servoing control law. In: IEEE Procs. of Int. Conference on Mechatronics and Automation (2014)
19. Millidge, B., Tschantz, A., Seth, A.K., Buckley, C.L.: On the relationship between active inference and control as inference. In: International Workshop on Active Inference. pp. 3–11. Springer (2020)
20. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
21. Narendra, K.S., Balakrishnan, J.: Adaptive control using multiple models. IEEE Trans. on Autom. Control (1997)
22. Oliver, G., Lanillos, P., Cheng, G.: An empirical study of active inference on a humanoid robot. IEEE Transactions on Cognitive and Developmental Systems pp. 1–1 (2021). https://doi.org/10.1109/TCDS.2021.3049907

23. Paviglianiti, G., Pierri, F., Caccavale, F., Mattei, M.: Robust fault detection and isolation for proprioceptive sensors of robot manipulators. Mechatronics **20(1)**, 162–170 (2010)
24. Pezzato, C., Ferrari, R., Corbato, C.H.: A novel adaptive controller for robot manipulators based on active inference. IEEE Robotics and Automation Letters **5**(2), 2973–2980 (2020)
25. Pezzato, C., Baioumy, M., Corbato, C.H., Hawes, N., Wisse, M., Ferrari, R.: Active inference for fault tolerant control of robot manipulators with sensory faults. In: Springer (ed.) 1st Int. Workshop on Active Inference, ECML PKDD. Communications in Computer and Information Science, vol. 1326 (2020)
26. Pfeifer, T., Lange, S., Protzel, P.: Dynamic covariance estimation—a parameter free approach to robust sensor fusion. In: 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). pp. 359–365. IEEE (2017)
27. Piepmeier, J., McMurray, G., Lipkin, H.: Uncalibrated dynamic visual servoing. In: IEEE Trans. on Robotics and Automation. vol. 20, pp. 143–147 (2004)
28. Pio-Lopez, L., Nizard, A., Friston, K., Pezzulo, G.: Active inference and robot control: a case study. Journal of The Royal Society Interface **13(122)** (2016)
29. Rostampour, V., Ferrari, R.M., Teixeira, A.M., Keviczky, T.: Privatized distributed anomaly detection for large-scale nonlinear uncertain systems. IEEE Trans. on Autom. Control (2020)
30. Shetty, A., Gao, G.X.: Covariance estimation for gps-lidar sensor fusion for uavs. In: Proceedings of the 30th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2017). pp. 2919–2923 (2017)
31. Van, M., Wu, D., Ge, S., Ren, H.: Fault diagnosis in image-based visual servoing with eye-in-hand configurations using Kalman filter. IEEE Trans. Industrial Electronics **12(6)**, 1998–2007 (2016)
32. Vega-Brown, W., Bachrach, A., Bry, A., Kelly, J., Roy, N.: Cello: A fast algorithm for covariance estimation. In: 2013 IEEE International Conference on Robotics and Automation. pp. 3160–3167. IEEE (2013)
33. Vega-Brown, W., Roy, N.: Cello-em: Adaptive sensor models without ground truth. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1907–1914. IEEE (2013)

On the Convergence of DEM's Linear Parameter Estimator

Ajith Anil Meera(✉) and Martijn Wisse

Cognitive Robotics, Delft Institute of Technology, The Netherlands
ajitham1994@gmail.com

Abstract. The free energy principle from neuroscience provides an efficient data-driven framework called the Dynamic Expectation Maximization (DEM), to learn the generative model in the environment. DEM's growing potential to be the brain-inspired learning algorithm for robots demands a mathematically rigorous analysis using the standard control system tools. Therefore, this paper derives the mathematical proof of convergence for its parameter estimator for linear state space systems, subjected to colored noise. We show that the free energy based parameter learning converges to a stable solution for linear systems. The paper concludes by providing a proof of concept through simulation for a wide range of spring damper systems.

Keywords: Free energy principle · Dynamic Expectation Maximization · Parameter estimation · Linear state space systems.

1 Introduction

The free energy principle (FEP) models the brain's perception and action as a gradient ascend over its free energy objective [7]. The action side of FEP, known as active inference [8], has already been applied to real robots including ground robots for SLAM [5], humanoid robots for body perception [12] and manipulator robots for pick and place operation [13]. Similarities with standard control technique like PID was also analyzed [3]. One of the variants of FEP, the Dynamic Expectation Maximization (DEM) [9], provides a model inversion framework for perception and system identification. DEM's distinctive feature lies in its capability to gracefully handle colored noise through the use of generalized coordinates [6], thereby rendering it with the potential to be the learning algorithm for robots. DEM was reformulated as a linear state and input observer under colored noise [11] and was validated for quadrotor flights [4]. A DEM based linear parameter estimator for system identification was developed by [2] and was applied for the perception of quadrotor in wind [1]. Since an estimator with convergence guarantees is preferred for safe robotics applications, we aim at paving way to DEM's practical use by mathematically analyzing it for its convergence properties. Moreover, it is of interest to the active inference community to develop active learning and control strategies with stability guarantees. The presence of generalized coordinates, mean field terms and brain

priors complicates the convergence proof and makes it different from other estimators like Expectation Maximization [10]. The goal of this paper is: 1) to show that DEM has convergence guarantees for linear systems with colored noise, and 2) to show that it can be applied to control system problems like the estimation of a mass-spring-damper system.

2 Preliminaries

Consider the linear plant dynamics (generative process) given in Eq. 1, where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are constant system matrices, $\mathbf{x} \in \mathbb{R}^n$ is the hidden state, $\mathbf{v} \in \mathbb{R}^r$ is the input and $\mathbf{y} \in \mathbb{R}^m$ is the output.

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{v} + \mathbf{w}, \qquad \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{z}. \tag{1}$$

Here $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m$ represent the process and measurement noise respectively. The notations of the plant are denoted in boldface, whereas its estimates are denoted in nonboldface letters. Since the brain has no access to the plant dynamics except through the sensory measurements $\mathbf{y}$, it maintains the copy of an approximate model of the generative process called the generative model. The noise color assumption (convolution of white noise with a Gaussian kernel) facilitates the differentiated form of the generative model as [9]:

$$\begin{aligned} x' &= Ax + Bv + w & y &= Cx + z \\ x'' &= Ax' + Bv' + w' & \dot{y} &= Cx' + z' \\ &... & &... \end{aligned} \tag{2}$$

One of the key technique behind DEM to model the colored noise is to express the time varying components in generalized coordinates, denoted by a tilde operator. The colored noises can be expressed in generalized coordinates using their higher derivatives as $\tilde{z} = [z, z', z'', ...]^T$ and $\tilde{w} = [w, w', w'', ...]^T$. The generative model in Eq. 2 can be compactly written as [9]:

$$\dot{\tilde{x}} = D^x\tilde{x} = \tilde{A}\tilde{x} + \tilde{B}\tilde{v} + \tilde{w} \qquad \tilde{y} = \tilde{C}\tilde{x} + \tilde{z} \tag{3}$$

where $D^x = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}_{(p+1)\times(p+1)} \otimes I_{n\times n}$, $\tilde{A} = I_{p+1} \otimes A$, $\tilde{B} = I_{p+1} \otimes B$ and $\tilde{C} = I_{p+1} \otimes C$. Here $\otimes$ is the Kronecker tensor product. To facilitate the convergence proof later in the paper, we introduce a redefinition for Eq. 3 with all parameters grouped to the right side as θ:

$$\dot{\tilde{x}} = M\theta + \tilde{w},\ \tilde{y} = N\theta + \tilde{z},\ \theta = \begin{bmatrix} vec(A^T) \\ vec(B^T) \\ vec(C^T) \end{bmatrix}, \tag{4}$$

where

$$M = \begin{bmatrix} I_n \otimes x^T & I_n \otimes v^T & I_n \otimes O_{1\times m} \\ I_n \otimes x'^T & I_n \otimes v'^T & I_n \otimes O_{1\times m} \\ ... & ... & ... \end{bmatrix}, N = \begin{bmatrix} I_n \otimes O_{1\times n} & I_n \otimes O_{1\times r} & I_m \otimes x^T \\ I_n \otimes O_{1\times n} & I_n \otimes O_{1\times r} & I_m \otimes x'^T \\ ... & ... & ... \end{bmatrix}. \tag{5}$$

The goal of this paper is to mathematically prove that the DEM's estimate for θ converges while maximizing the free energy objective.

3 Parameter learning as free energy optimization

DEM postulates the parameter learning algorithm as the gradient ascend over the free energy action, which is the time integral of free energy $\bar{F} = \int F dt$. The parameter update equation can be expressed as the gradient [9, 2]:

$$\frac{\partial \theta}{\partial a} = k^\theta \frac{\partial \bar{F}}{\partial \theta} = -P^\theta(\theta - \eta^\theta) + \sum_t (-E_\theta + W_\theta^X), \tag{6}$$

where k^θ is the learning rate, $E_\theta = \frac{\partial E}{\partial \theta}$ is the gradient of precision weighed prediction error, $W_{\theta^i}^X = \frac{\partial W^X}{\partial \theta}$ is the gradient of state mean field term, η^θ is the prior parameters and P^θ is the prior parameter precision. Subscripts will be used for the derivative operator. E_θ for an LTI system can be simplified as:

$$E_\theta = \tilde{\epsilon}_\theta^T \tilde{\Pi} \tilde{\epsilon}, \text{ where } \tilde{\epsilon} = \begin{bmatrix} \tilde{\mathbf{y}} - N\theta \\ \tilde{v} - \tilde{\eta}^v \\ D^x \tilde{x} - M\theta \end{bmatrix} \text{ and } \tilde{\epsilon}_\theta = \begin{bmatrix} -N \\ O \\ -M \end{bmatrix} \tag{7}$$

are the prediction error and its gradient. Here $\tilde{\eta}^v$ is the prior on inputs with prior precision $\tilde{P}^v$, $\tilde{\Pi} = diag(\tilde{\Pi}^z, \tilde{P}^v, \tilde{\Pi}^w)$ is the generalized noise precision with Π^z and Π^w being the precisions (inverse covariance) for measurement and process noise. Here $diag()$ represents the block diagonal operation. Similarly, W_θ^X for an LTI system can be written as [9, 2]:

$$W_{\theta^i}^X = -\frac{1}{2} tr(\Sigma^X \tilde{\epsilon}_{X\theta^i}^T \tilde{\Pi} \tilde{\epsilon}_X), \ \tilde{\epsilon} = \begin{bmatrix} \tilde{\mathbf{y}} - \tilde{C}\tilde{x} \\ \tilde{v} - \tilde{\eta}^v \\ D^x \tilde{x} - \tilde{A}\tilde{x} - \tilde{B}\tilde{v} \end{bmatrix}, \ \tilde{\epsilon}_X = \begin{bmatrix} -\tilde{C} & O \\ O & I \\ D^x - \tilde{A} & -\tilde{B} \end{bmatrix}. \tag{8}$$

4 Proof of convergence for parameter estimator

If E_θ and W_θ^X can be expressed as linear in θ, in the form $E_\theta = E_1\theta + E_2$ and $W_\theta^X = W_1\theta + W_2$, Eq. 6 can be rewritten as:

$$\frac{\partial \theta}{\partial a} = -\Big[P^\theta + \sum_t (E_1 - W_1)\Big]\theta + \Big[P^\theta \eta^\theta + \sum_t (-E_2 + W_2)\Big]. \tag{9}$$

The differential equation given by Eq. 9 is of the form of a linear state space equation ($\dot{\theta} = A^{\theta}\theta + B^{\theta}.1$). From the basics of control theory, the solutions of this equation converges exponentially (stabilise) if $A^{\theta} = -[P^{\theta} + \sum_t (E_1 - W_1)]$ is negative definite (negative eigen values). This section aims to prove this result.

Lemma 1. *If $A, B \succ O$, then $A + B \succ O$.*

As per Lemma 1, the positive definiteness of $P^{\theta} - \sum_t W_1 + \sum_t E_1$ can be proved by proving the positive definiteness of the individual terms P^{θ}, $-W_1$ and E_1. We know by definition that the prior parameter precision P^{θ} is positive definite. We now proceed to prove that $E_1 \succeq O$. Upon simplification of Eq. 7, E_{θ} can be written as $E_{\theta} = E_1\theta + E_2$, where:

$$E_1 = N^T \tilde{\Pi}^z N + M^T \tilde{\Pi}^w M \text{ and } E_2 = -\begin{bmatrix} N^T \tilde{\Pi}^z & M^T \tilde{\Pi}^w D \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{y}} \\ \tilde{x} \end{bmatrix}. \tag{10}$$

Lemma 2. *If $A \succeq O$, then $B^T A B \succeq O$.*

Proof. By definition, if $A \succeq O$, there exists a square root $A^{\frac{1}{2}} \succeq O$. Therefore, $x^T(B^T A B)x = x^T(B^T A^{\frac{1}{2}} A^{\frac{1}{2}} B)x = (A^{\frac{1}{2}} Bx)^T (A^{\frac{1}{2}} Bx) \geq 0, \implies B^T A B \succeq O$.

Since $\tilde{\Pi}^z \succ O$ and $\tilde{\Pi}^w \succ O$ by definition, from Lemma 1 and 2, $E_1 = N^T \tilde{\Pi}^z N + M^T \tilde{\Pi}^w M \succeq O$. Therefore, E_1 is proved to be positive semi-definite.

The final term under consideration is W_1. The rest of this section aims to prove that $W_1 \prec O$, which will conclude the entire convergence proof of parameter estimation. We rewrite the mean field term for parameter θ^i from Eq. 8 as:

$$\begin{aligned} W^X_{\theta^i} &= -\frac{1}{2} tr(\Sigma^X \tilde{\epsilon}^T_{X\theta^i} \tilde{\Pi} \tilde{\epsilon}_X), \\ &= -\frac{1}{2} tr\left[\begin{bmatrix} \Sigma^{\tilde{x}\tilde{x}} & \Sigma^{\tilde{x}\tilde{v}} \\ \Sigma^{\tilde{v}\tilde{x}} & \Sigma^{\tilde{v}\tilde{v}} \end{bmatrix} \begin{bmatrix} \tilde{C}^T_{\theta^i} \tilde{\Pi}^z \tilde{C} - \tilde{A}^T_{\theta^i} \tilde{\Pi}^w (D - \tilde{A}) & \tilde{A}^T_{\theta^i} \tilde{\Pi}^w \tilde{B} \\ -\tilde{B}^T_{\theta^i} \tilde{\Pi}^w (D - \tilde{A}) & \tilde{B}^T_{\theta^i} \tilde{\Pi}^w \tilde{B} \end{bmatrix} \right] \\ &= -\frac{1}{2} tr\left[\begin{bmatrix} \Sigma^{\tilde{x}\tilde{x}} & \Sigma^{\tilde{x}\tilde{v}} \\ \Sigma^{\tilde{v}\tilde{x}} & \Sigma^{\tilde{v}\tilde{v}} \end{bmatrix} \begin{bmatrix} \tilde{C}^T_{\theta^i} \tilde{\Pi}^z \tilde{C} + \tilde{A}^T_{\theta^i} \tilde{\Pi}^w \tilde{A} & \tilde{A}^T_{\theta^i} \tilde{\Pi}^w \tilde{B} \\ \tilde{B}^T_{\theta^i} \tilde{\Pi}^w \tilde{A} & \tilde{B}^T_{\theta^i} \tilde{\Pi}^w \tilde{B} \end{bmatrix} \right] \\ &\quad -\frac{1}{2} tr\left[\begin{bmatrix} \Sigma^{\tilde{x}\tilde{x}} & \Sigma^{\tilde{x}\tilde{v}} \\ \Sigma^{\tilde{v}\tilde{x}} & \Sigma^{\tilde{v}\tilde{v}} \end{bmatrix} \begin{bmatrix} -\tilde{A}^T_{\theta^i} \tilde{\Pi}^w D & O \\ -\tilde{B}^T_{\theta^i} \tilde{\Pi}^w D & O \end{bmatrix} \right]. \end{aligned} \tag{11}$$

Since the second trace term in Eq. 11 is independent of θ^i, it is lumped into the $W_2^{\theta^i}$ term. Equation 11 is further simplified as:

$$\begin{aligned} W^X_{\theta^i} = -\frac{1}{2}\Big[& tr(\Sigma^{\tilde{x}\tilde{x}} \tilde{C}^T_{\theta^i} \tilde{\Pi}^z \tilde{C}) + tr(\Sigma^{\tilde{x}\tilde{x}} \tilde{A}^T_{\theta^i} \tilde{\Pi}^w \tilde{A}) + tr(\Sigma^{\tilde{x}\tilde{v}} \tilde{B}^T_{\theta^i} \tilde{\Pi}^w \tilde{A}) \\ & + tr(\Sigma^{\tilde{v}\tilde{x}} \tilde{A}^T_{\theta^i} \tilde{\Pi}^w \tilde{B}) + tr(\Sigma^{\tilde{v}\tilde{v}} \tilde{B}^T_{\theta^i} \tilde{\Pi}^w \tilde{B}) \Big] + W_2^{\theta^i} \end{aligned} \tag{12}$$

We aim to separate θ out so that the mean field term can be expressed in the form $W^X_{\theta} = W_1\theta + W_2$. We proceed by first introducing the transpose of the generalized parameter matrices $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ to Eq. 12 and then moving them out of the trace terms.

Lemma 3. *If A, B, C and D are matrices, then* $tr(ABCD) = tr(C^T B^T A^T D^T)$

Proof. $tr(ABCD) = tr((ABCD)^T) = tr(D^T C^T B^T A^T) = tr(C^T B^T A^T D^T)$.

Lemma 4. *If A, B and C are matrices, then* $tr(ABC) = vec(A^T)^T (I \otimes B) vec(C)$.

Applying Lemma 3 throughout Eq. 12 results in:

$$\begin{aligned} W_{\theta^i}^X = -\frac{1}{2}\Big[& tr(\tilde{\Pi}^{zT} \tilde{C_{\theta^i}} \Sigma^{\tilde{x}\tilde{x}T} \tilde{C}^T) + tr(\tilde{\Pi}^{wT} \tilde{A_{\theta^i}} \Sigma^{\tilde{x}\tilde{x}T} \tilde{A}^T) + tr(\tilde{\Pi}^{wT} \tilde{B_{\theta^i}} \Sigma^{\tilde{x}\tilde{v}T} \tilde{A}^T) \\ & + tr(\tilde{\Pi}^{wT} \tilde{A_{\theta^i}} \Sigma^{\tilde{v}\tilde{x}T} \tilde{B}^T) + tr(\tilde{\Pi}^{wT} \tilde{B_{\theta^i}} \Sigma^{\tilde{v}\tilde{v}T} \tilde{B}^T)\Big] + W_2^{\theta^i}, \end{aligned} \tag{13}$$

which upon further expansion using Lemma 4 and grouping yields:

$$\begin{aligned} W_{\theta^i}^X = -\frac{1}{2}\Bigg[& \Big(vec(\tilde{A}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{x}\tilde{x}T}) + vec(\tilde{B}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{x}\tilde{v}T})\Big) vec(\tilde{A}^T) \\ & + \Big(vec(\tilde{A}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{v}\tilde{x}T}) + vec(\tilde{B}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{v}\tilde{v}T})\Big) vec(\tilde{B}^T) \\ & + \Big(vec(\tilde{C}_{\theta^i}^T \tilde{\Pi}^z)^T (I \otimes \Sigma^{\tilde{x}\tilde{x}T})\Big) vec(\tilde{C}^T)\Bigg] + W_2^{\theta^i}. \end{aligned} \tag{14}$$

We have now separated all the generalized parameters out of the trace terms in their vector forms. These vectors can be grouped such that the mean field term is linear with respect to the generalized parameter vector $\tilde{\theta} = \begin{bmatrix} vec(\tilde{A}^T) \\ vec(\tilde{B}^T) \\ vec(\tilde{C}^T) \end{bmatrix}$ as:

$$\begin{aligned} W_{\theta^i}^X = -\frac{1}{2}\Bigg[& vec(\tilde{A}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{x}\tilde{x}T}) + vec(\tilde{B}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{x}\tilde{v}T}), \\ & vec(\tilde{A}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{v}\tilde{x}T}) + vec(\tilde{B}_{\theta^i}^T \tilde{\Pi}^w)^T (I \otimes \Sigma^{\tilde{v}\tilde{v}T}), \\ & vec(\tilde{C}_{\theta^i}^T \tilde{\Pi}^z)^T (I \otimes \Sigma^{\tilde{x}\tilde{x}T})\Bigg] \tilde{\theta} + W_2^{\theta^i}. \end{aligned} \tag{15}$$

Lemma 5. *If A and B are matrices, then* $vec(AB)^T = vec(A)^T (B \otimes I)$.

We use Lemma 5 to further simplify Eq. 15 as:

$$\begin{aligned} W_{\theta^i}^X = -\frac{1}{2}\Bigg[& vec(\tilde{A}_{\theta^i}^T)^T (\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{x}\tilde{x}T}) + vec(\tilde{B}_{\theta^i}^T)^T (\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{x}\tilde{v}T}), \\ & vec(\tilde{A}_{\theta^i}^T)^T (\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{v}\tilde{x}T}) + vec(\tilde{B}_{\theta^i}^T)^T (\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{v}\tilde{v}T}), \\ & vec(\tilde{C}_{\theta^i}^T)^T (\tilde{\Pi}^z \otimes I)(I \otimes \Sigma^{\tilde{x}\tilde{x}T})\Bigg] \tilde{\theta} + W_2^{\theta^i}. \end{aligned} \tag{16}$$

Since the parameters A, B and C are independent of each other, their derivatives with respect to each other are zeros, resulting in $vec(\tilde{A}^T_{\theta^i}) = O, \forall \theta^i \in \{B, C\}$, $vec(\tilde{B}^T_{\theta^i}) = O, \forall \theta^i \in \{A, C\}$ and $vec(\tilde{C}^T_{\theta^i}) = O, \forall \theta^i \in \{A, B\}$. This simplifies the expression for $W^X_{\theta^i}$ in Eq. 16. The total mean field term W^X_θ can be computed by vertically stacking the individual mean field contributions $W^X_{\theta^i}$ from each parameter θ^i as:

$$W^X_\theta = -\frac{1}{2} W_3 \tilde{\theta} + W_2, \tag{17}$$

where $W_3 = \begin{bmatrix} W_4 & O \\ O & W_5, \end{bmatrix}$ with $W_5 = vec(\tilde{C}^T)^T_{vecC^T}(\tilde{\Pi}^z \otimes I)(I \otimes \Sigma^{\tilde{x}\tilde{x}T})$ and

$$W_4 = \begin{bmatrix} vec(\tilde{A}^T)^T_{vecA^T}(\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{x}\tilde{x}T}) & vec(\tilde{A}^T)^T_{vecA^T}(\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{v}\tilde{x}T}) \\ vec(\tilde{B}^T)^T_{vecB^T}(\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{x}\tilde{v}T}) & vec(\tilde{B}^T)^T_{vecB^T}(\tilde{\Pi}^w \otimes I)(I \otimes \Sigma^{\tilde{v}\tilde{v}T}) \end{bmatrix}.$$

W_3 can be simplified as:

$$W_3 = \frac{\partial \tilde{\theta}}{\partial \theta}^T \begin{bmatrix} \tilde{\Pi}^w \otimes I & O & O \\ O & \tilde{\Pi}^w \otimes I & O \\ O & O & \tilde{\Pi}^z \otimes I \end{bmatrix} \begin{bmatrix} I \otimes \Sigma^{\tilde{x}\tilde{x}T} & I \otimes \Sigma^{\tilde{v}\tilde{x}T} & O \\ I \otimes \Sigma^{\tilde{x}\tilde{v}T} & I \otimes \Sigma^{\tilde{v}\tilde{v}T} & O \\ O & O & I \otimes \Sigma^{\tilde{x}\tilde{x}T} \end{bmatrix}, \tag{18}$$

where $\frac{\partial \tilde{\theta}}{\partial \theta} = diag(vec\tilde{A}^T_{vecA^T}, vec\tilde{B}^T_{vecB^T}, vec\tilde{C}^T_{vecC^T})$. Since the generalized parameter vector $\tilde{\theta}$ is linear in parameter vector θ, we can write:

$$\tilde{\theta} = \frac{\partial \tilde{\theta}}{\partial \theta}\theta = \begin{bmatrix} vec\tilde{A}^T_{vecA^T} & O & O \\ O & vec\tilde{B}^T_{vecB^T} & O \\ O & O & vec\tilde{C}^T_{vecC^T} \end{bmatrix} \theta. \tag{19}$$

Substituting Eq. 18 and 19 in Eq. 17 yields:

$$W^X_\theta = W_1\theta + W_2,$$
$$W_1 = -\frac{1}{2}\frac{\partial \tilde{\theta}}{\partial \theta}^T \begin{bmatrix} \tilde{\Pi}^w \otimes I & O & O \\ O & \tilde{\Pi}^w \otimes I & O \\ O & O & \tilde{\Pi}^z \otimes I \end{bmatrix} \begin{bmatrix} I \otimes \Sigma^{\tilde{x}\tilde{x}T} & I \otimes \Sigma^{\tilde{v}\tilde{x}T} & O \\ I \otimes \Sigma^{\tilde{x}\tilde{v}T} & I \otimes \Sigma^{\tilde{v}\tilde{v}T} & O \\ O & O & I \otimes \Sigma^{\tilde{x}\tilde{x}T} \end{bmatrix} \frac{\partial \tilde{\theta}}{\partial \theta}. \tag{20}$$

Therefore, the mean field term W^X_θ is linear in θ. For the parameter estimator to provide a converging solution, we need to prove that $W_1 \prec O$. Lemma 2 could be applied to the expression for W_1 to prove that $W_1 \prec O$ if:

$$W_6 = \begin{bmatrix} \tilde{\Pi}^w \otimes I & O & O \\ O & \tilde{\Pi}^w \otimes I & O \\ O & O & \tilde{\Pi}^z \otimes I \end{bmatrix} \begin{bmatrix} I \otimes \Sigma^{\tilde{x}\tilde{x}T} & I \otimes \Sigma^{\tilde{v}\tilde{x}T} & O \\ I \otimes \Sigma^{\tilde{x}\tilde{v}T} & I \otimes \Sigma^{\tilde{v}\tilde{v}T} & O \\ O & O & I \otimes \Sigma^{\tilde{x}\tilde{x}T} \end{bmatrix} \succ O \tag{21}$$

Lemma 6. *If $A, B \succeq O$ and A is invertible, then $AB \succeq O$.*

Proof. $AB = A^{\frac{1}{2}}(A^{\frac{1}{2}}BA^{\frac{1}{2}})A^{-\frac{1}{2}}$, implies AB and $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ are similar matrices, sharing all eigen values. Using lemma 2, since $B \succeq O$, $A^{\frac{1}{2}}BA^{\frac{1}{2}} \succeq O \implies AB \succeq O$.

Using lemma 6 it is straightforward to see that $W_6 \succeq O$ because: $\tilde{\Pi}^z \succ O, \tilde{\Pi}^w \succ O, \implies \tilde{\Pi}^z \otimes I \succ O$ and $\tilde{\Pi}^w \otimes I \succ O, I \otimes \Sigma^X \succ O$. Therefore, $W_1 \preceq O$. This completes the proof that the parameter estimation of DEM converges for an LTI system with colored noise.

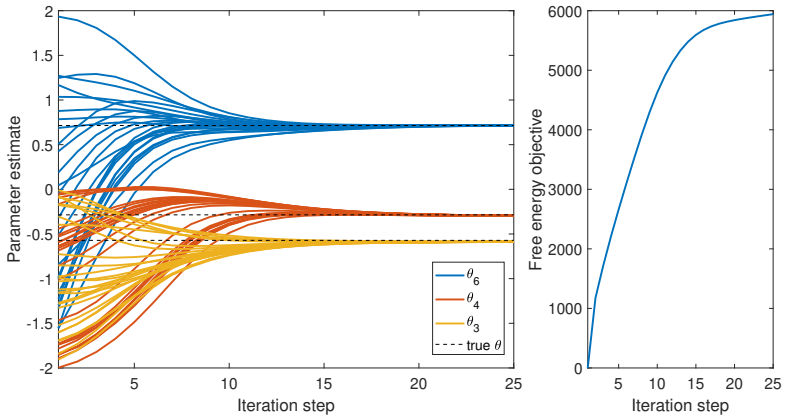


Fig. 1. The parameter estimates of DEM converges to the correct value of $\theta_3 = -\frac{k}{m} = -0.5714$, $\theta_4 = -\frac{b}{m} = -0.2857$ and $\theta_6 = \frac{1}{m} = 0.7143$, marked in black, for a set of 25 experiments, despite being initialized by randomly sampled priors such that $\eta^{\theta^i} \in [-2, 2]$ and that the prior A matrix is stable. The parameter estimation proceeds by maximizing the free energy objective as shown on the right (sample realization).

5 Proof of concept: mass-spring-damper system

This section aims at providing a proof of concept for the convergence of DEM's parameter estimator, through realistic simulations. A mass-spring-damper system with mass $m = 1.4kg$, spring constant $k = 0.8N/m$ and damping coefficient $b = 0.4Ns/m$, is considered in the state space form given by:

$$\begin{bmatrix} \dot{x} \\ \ddot{x} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} v, \; y = \begin{bmatrix} 1 \; 0 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \end{bmatrix}. \tag{22}$$

A Gaussian bump input $v = e^{-0.25(t-12)^2}$, centred around 12s and sampled at $dt = 0.1s$ for $T = 32s$ was used. To generate the colored noise, the white noise ($\Pi^w = e^6 I_2$ and $\Pi^z = e^6$) was convoluted using a Gaussian kernel with a width

of $\sigma = 0.5s$. A partially known system with unknown $\theta_3 = -\frac{k}{m}$, $\theta_4 = -\frac{b}{m}$ and $\theta_6 = \frac{1}{m}$ was considered. Using the output $\mathbf{y}$ generated from the spring damper system, parameter estimation was performed using DEM for 25 experiments with different η^θ. The parameter priors η^θ for unknown parameters were randomly sampled from [-2,2] such that the resulting prior A matrix is stable. A low prior precision ($P^{\theta_i} = e^{-4}$) was used for known parameters, and a high precision ($P^{\theta_i} = e^{32}$) was used for unknown parameters. The order of generalized motion of $p = 6$ and $d = 2$ were used for the states and inputs respectively. The result for DEM's parameter estimation is shown in Fig. 1. Despite being initialized by random wrong priors, DEM's parameter estimates exponentially converges to the correct values, by maximizing the free energy objective.

Next, we proceed to show that the estimate converges for a wide range of systems. The same experiment was repeated for 25 different randomly selected stable mass-spring-damper systems. Although the convergence applies to unstable systems, sampling was restricted to stable systems within the range [-1,1] ($\theta_3, \theta_4 \in [-1, 0]$ and $\theta_6 \in [0, 1]$) for better visualization. DEM was initialized with the same priors for all experiments ($\eta^{\theta_6} = 2$, $\eta^{\theta_4} = -1$ and $\eta^{\theta_3} = -2$). Figure 2 shows that DEM is capable of providing converging solutions for a wide range of stable spring-damper systems, that are influenced by colored noise. Note that the numerical analysis is restricted to the dynamics of spring damper systems for demonstrative purposes, and can be extended to other systems. In summary, DEM can provide converging parameter estimates for linear systems with colored noise, by maximizing the free energy objective.

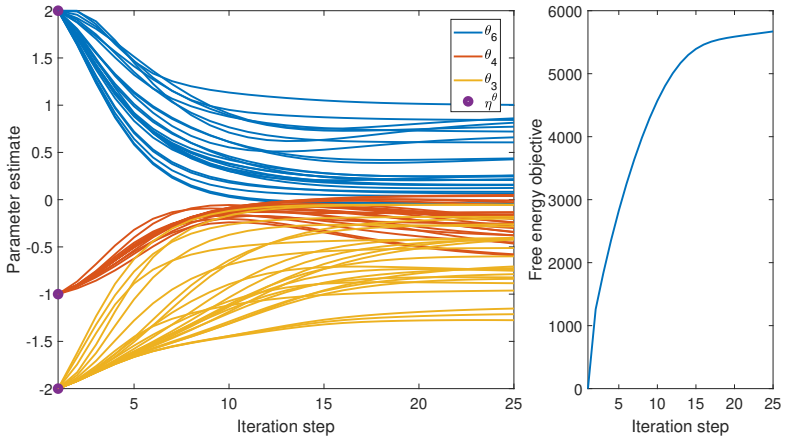


Fig. 2. DEM's parameter estimates for 25 different randomly sampled stable mass-spring-damper systems. The estimates for all the experiments started from the same prior of $\eta^{\theta_6} = 2$, $\eta^{\theta_4} = -1$ and $\eta^{\theta_3} = -2$, and converged, while maximizing the free energy objective. Therefore, the estimator converges for a wide range of systems.

6 Conclusion

DEM has the potential to be a bioinspired learning algorithm for future robots, due to its capability to robustly handle colored noise. Its superior performance in state estimation under colored noise was proven by [11] and was experimentally validated by [4]. In this paper, we derived a mathematical proof of convergence for DEM's parameter estimator, applied to linear systems with colored noise. We proved that a perception scheme based on the gradient ascend of the free energy action, provides a converging solution. Since a convergence proof is mandatory for the safe and reliable application of DEM on real robots, this work widens its applicability in robotics. The applicability of DEM for real control system problem was demonstrated through rigorous simulations on the estimation problem for mass-spring-damper systems. The future research will focus on the conditions for unbiased estimation and on applying DEM to real robots.

References

1. Anil Meera, A., Wisse, M.: A brain inspired learning algorithm for the perception of a quadrotor in wind, *Under review* (2021)
2. Anil Meera, A., Wisse, M.: Dynamic expectation maximization algorithm for estimation of linear systems with colored noise, *Under review* (2021)
3. Baltieri, M., Buckley, C.L.: Pid control as a process of active inference with linear generative models. Entropy **21**(3), 257 (2019)
4. Bos, F., Anil Meera, A., Benders, D., Wisse, M.: Free energy principle for state and input estimation of a quadcopter flying in wind, *Under review* (2021)
5. Çatal, O., Verbelen, T., Van de Maele, T., Dhoedt, B., Safron, A.: Robot navigation as hierarchical active inference. Neural Networks **142**, 192–204 (2021)
6. Friston, K.: Hierarchical models in the brain. PLoS computational biology **4**(11), e1000211 (2008)
7. Friston, K.: The free-energy principle: a unified brain theory? Nature reviews neuroscience **11**(2), 127–138 (2010)
8. Friston, K., Mattout, J., Kilner, J.: Action understanding and active inference. Biological cybernetics **104**(1), 137–160 (2011)
9. Friston, K.J., Trujillo-Barreto, N., Daunizeau, J.: DEM: a variational treatment of dynamic systems. Neuroimage **41**(3), 849–885 (2008)
10. Mader, W., Linke, Y., Mader, M., Sommerlade, L., Timmer, J., Schelter, B.: A numerically efficient implementation of the expectation maximization algorithm for state space models. Applied Mathematics and Computation **241**, 222–232 (2014)
11. Meera, A.A., Wisse, M.: Free energy principle based state and input observer design for linear systems with colored noise. In: 2020 American Control Conference (ACC). pp. 5052–5058. IEEE (2020)
12. Oliver, G., Lanillos, P., Cheng, G.: Active inference body perception and action for humanoid robots. arXiv preprint arXiv:1906.03022 (2019)
13. Pezzato, C., Ferrari, R., Corbato, C.H.: A novel adaptive controller for robot manipulators based on active inference. IEEE Robotics and Automation Letters **5**(2), 2973–2980 (2020)

Disentangling What and Where for 3D Object-Centric Representations Through Active Inference

Toon Van de Maele, Tim Verbelen, Ozan Çatal and Bart Dhoedt

IDLab, Department of Information Technology
Ghent University - imec
Ghent, Belgium
firstname.lastname@ugent.be

Abstract. Although modern object detection and classification models achieve high accuracy, these are typically constrained in advance on a fixed train set and are therefore not flexible to deal with novel, unseen object categories. Moreover, these models most often operate on a single frame, which may yield incorrect classifications in case of ambiguous viewpoints. In this paper, we propose an active inference agent that actively gathers evidence for object classifications, and can learn novel object categories over time. Drawing inspiration from the human brain, we build object-centric generative models composed of two information streams, a what- and a where-stream. The what-stream predicts whether the observed object belongs to a specific category, while the where-stream is responsible for representing the object in its internal 3D reference frame. We show that our agent (i) is able to learn representations for many object categories in an unsupervised way, (ii) achieves state-of-the-art classification accuracies, actively resolving ambiguity when required and (iii) identifies novel object categories. Furthermore, we validate our system in an end-to-end fashion where the agent is able to search for an object at a given pose from a pixel-based rendering. We believe that this is a first step towards building modular, intelligent systems that can be used for a wide range of tasks involving three dimensional objects.

Keywords: Deep Learning, Object Recognition, Object Pose Estimation, Active Inference

1 Introduction

In the last decade, we have seen a proliferation of deep learning systems, especially in the field of image classificaton [15, 10]. Although these systems show high accuracies on various classification benchmarks, their applicability is typically limited to a fixed input distribution based on the dataset used during training. In contrast, the real world is not stationary, which urges the need for continual learning [7]. Also, these classifiers lack the concept of action, which renders them vulnerable to ambiguous and adverserial samples [6]. As humans,

we will typically move around and sample more viewpoints to improve the precision of our classification, illustrating the importance of embodiment in building intelligent agents [22].

Active inference offers a unified treatment of perception, action and learning, which states that intelligent systems build a generative model of their world and operate by minimizing a bound on surprise, i.e. the variational free energy [5]. In [18], Parr et al. propose a model for (human) vision, which considers a scene as a factorization of separate (parts of) objects, encoding their identity, scale and pose. This is in line with the so called two stream hypothesis, which states that visual information is processed by a dorsal ("where") stream on the one hand, representing where an object is in the space, and a ventral ("what") stream on the other hand, representing object identity [4]. Similarly, Hawkins et al. propose that cortical columns in the neocortex track objects and their pose in a local reference frame, encoded by cortical grid cells [9].

In this paper, we propose a system that builds on these principles for learning object-centric representations that allow for accurate classification. Inspired by cortical columns, our system is composed of separate deep neural networks, called Cortical Column Networks (CCN), where each CCN learns a representation of a single type of 3D object in a local reference frame. The ensemble of CCNs forms the agent's generative model, which is optimized by minimizing free energy. By also minimizing the expected free energy in the future, we show that our agent can realize preferred viewpoints for certain objects, while also being urged to resolve ambiguity on object identiy.

We evaluate our agent on pixel data rendered from 3D objects from the YCB benchmarking dataset [1], where the agent can control the viewpoint. We compare the performance of an embodied and a static agent for classification, and show that classification accuracy is higher for the embodied agent. Additionally, we leverage the where stream for implicit pose estimation of the objects.

2 Method

In active inference, an agent acts and learns in order to minimize an upper bound on the negative log evidence of its observations, given its generative model of the world i.e. the free energy. In this section, we first formally introduce the generative model of our agent for representing 3D objects. Next we discuss how we instantiate and train this generative model using deep neural networks. Finally, we show how action selection is driven by minimizing expected free energy in the future.

2.1 A generative model for object-centric perception

Our generative model is based on [18], but focused on representing a single object. Concretely, our agent obtains pixel observations $\mathbf{o}_{0:t}$ that render a 3D object with identity $\mathbf{i}$ as viewed from certain viewpoints $\mathbf{v}_{0:t}$ specified in an object-local

reference frame. Each time step t the agent can perform an action $\mathbf{a}_t$, resulting in a relative translation and rotation of the camera. The joint probability distribution then factorizes as:

$$p(\mathbf{o}_{0:t}, \mathbf{a}_{0:t-1}, \mathbf{v}_{0:t}, \mathbf{i}) = p(\mathbf{i}) \prod_t p(\mathbf{o}_t|\mathbf{v}_t, \mathbf{i}) p(\mathbf{v}_t|\mathbf{v}_{t-1}, \mathbf{a}_{t-1}) p(\mathbf{a}_{t-1}) \tag{1}$$

Using the approximate posterior $q(\mathbf{i}, \mathbf{v}_{0:t}|\mathbf{o}_{0:t}) = q(\mathbf{i}|\mathbf{o}_{0:t}) \sum_t q(\mathbf{v}_t|\mathbf{i}, \mathbf{o}_t)$, the free energy becomes:

$$\begin{aligned} F &= \mathbb{E}_{q(\mathbf{i}, \mathbf{v}_{0:t})}[\log q(\mathbf{i}, \mathbf{v}_{0:t}|\mathbf{o}_{0:t}) - \log p(\mathbf{o}_{0:t}, \mathbf{a}_{0:t-1}, \mathbf{v}_{0:t}, \mathbf{i})] \\ &\stackrel{+}{=} D_{KL}[q(\mathbf{i}|\mathbf{o}_{0:t})||p(\mathbf{i})] + \sum_t D_{KL}[q(\mathbf{v}_t|\mathbf{i}, \mathbf{o}_t)||p(\mathbf{v}_t|\mathbf{v}_{t-1}, \mathbf{a}_{t-1})] \\ &\quad - \mathbb{E}_{q(\mathbf{i}, \mathbf{v}_{0:t})}[\log p(\mathbf{o}_t|\mathbf{v}_t, \mathbf{i})] \end{aligned} \tag{2}$$

This shows that minimizing free energy is equivalent to maximizing the accuracy, i.e. predicting the observation for a given object identity and viewpoint, while minimizing complexity of the posterior models.

2.2 An ensemble of CCNs

We instantiate the generative model using deep neural networks similar to a variational autoencoder (VAE) [14, 19] with an encoder and decoder part. For each object identity, we train a separate encoder-decoder pair, since $p(\mathbf{o}_t|\mathbf{v}_t, \mathbf{i}) = \sum_k p(\mathbf{o}_t|\mathbf{v}_t, i = k)$. Similarly the encoder outputs distribution parameters for the object identity $q(i = k|\mathbf{o}_t)$ and viewpoint $q(\mathbf{v}_t|\mathbf{o}_t, i = k)$, the former parameterized as a Bernoulli variable, the latter as a multivariate Gaussian with a diagonal covariance matrix. Finally, we also parameterize the transition model $p(\mathbf{v}_t|\mathbf{v}_{t-1}, \mathbf{a}_{t-1}, i = k)$ which enforces $\mathbf{v}$ to encode relative viewpoint information.

Intuitively, each encoder-decoder pair captures the information about a single object class, with a "what" stream modeled as a binary classifier of whether an observation belongs to a certain object identiy, and a "where" stream encoding the observer viewpoint w.r.t. a local, object-specific reference frame. We call such a pair a Cortical Column Network (CCN), as it mimicks the "voting for object at pose" behavior of cortical columns in the neocortex as hypothesized in [9]. This is illustrated in Figure 1. The agent hence entails a generative model as an ensemble of CCNs. We obtain $q(\mathbf{i}|\mathbf{o}_{0:t}) \propto q(\mathbf{i}|\mathbf{o}_{0:t-1}) q(\mathbf{i}|\mathbf{o}_t)$, where $q(\mathbf{i}|\mathbf{o}_t)$ is a Categorical distribution from the CCN votes $q(i = k|\mathbf{o}_t)$, and $q(\mathbf{i}|\mathbf{o}_{0:t-1})$ is a conjugate prior Dirichlet distribution whose concencentration parameters are aggregated votes from previous observations, as updated in a Bayesian filter [25]. This process computes the posterior belief over the different timesteps. The Dirichlet distribution reflects the prior that an object is unlikely to change its category between timesteps. We also include an "other" object class, which is activate when none of the object classes receive votes, hence enabling the agent to detect novel object categories.

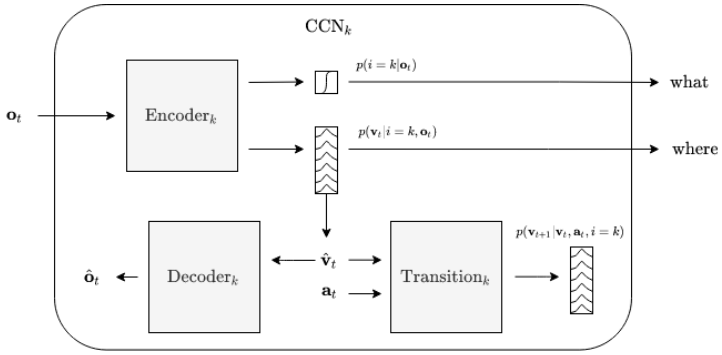


Fig. 1: A single CCN. Observation $\mathbf{o}_t$ is processed by an encoder and provides a belief over the object identity $p(i = k|\mathbf{o}_t)$ and over the observers pose $p(\mathbf{v}_t|i = k, \mathbf{o}_t)$. From this distribution, a sample $\hat{\mathbf{v}}_t$ is drawn and is decoded in a reconstructed observation $\hat{\mathbf{o}}_t$. This sample is also transformed together with an action $\mathbf{a}_t$ into a belief over a future pose $\mathbf{v}_{t+1}$.

Each CCN is trained in an end-to-end fashion using a dataset of object observation pairs and the relative camera transform between them for each object class. To minimize Equation 2, we use MSE loss on the reconstructions and a KL divergence between the viewpoint posterior and transition model. The identity posterior is trained as a binary classifier, sampling positive and negative anchors from the dataset. For more details on the training loss and model architectures, the reader is referred to the appendix.

2.3 Classification by minimizing expected free energy

Crucially in active inference, an agent will select the action that minimizes the expected free energy in the future G. In our case, this yields:

$$\begin{aligned} G(\mathbf{a}_t) &= \mathbb{E}_{q(\mathbf{i},\mathbf{v}_{0:t+1},\mathbf{o}_{t+1})}[\log q(\mathbf{i}, \mathbf{v}_{0:t+1}|\mathbf{o}_{0:t}, \mathbf{a}_t) - \log p(\mathbf{o}_{0:t+1}, \mathbf{a}_{0:t-1}, \mathbf{v}_{0:t+1}, \mathbf{i}|\mathbf{a}_t)] \\ &\approx \mathbb{E}_{q(\mathbf{o}_{t+1})}[-\log p(\mathbf{o}_{0:t+1})] \\ &\quad - \mathbb{E}_{q(\mathbf{i},\mathbf{v}_{0:t+1},\mathbf{o}_{t+1})}[\log q(\mathbf{i}|\mathbf{o}_{0:t+1}, \mathbf{a}_t) - \log q(\mathbf{i}|\mathbf{o}_{0:t}, \mathbf{a}_t)] \\ &\quad - \mathbb{E}_{q(\mathbf{i},\mathbf{v}_{0:t+1},\mathbf{o}_{t+1})}[\log q(\mathbf{v}_{0:t+1}|\mathbf{i}, \mathbf{o}_{0:t+1}, \mathbf{a}_t) - \log q(\mathbf{v}_{0:t+1}|\mathbf{i}, \mathbf{o}_{0:t}, \mathbf{a}_t)] \quad (3) \end{aligned}$$

The expected free energy unpacks into three terms, the first is an instrumental term that indicates that the agent is driven to some prior preferred observations, whereas the second and third term encode the expected information gain for the object identity and the object pose for a certain action. This shows how the agent can be steered to seeing a certain object at a certain pose, which could be for example a grasp position in the case of a robotic manipulator. On the other hand, in the absence of preferences, the agent will query new viewpoints that provide information on the object identity and pose, effectively trying to get a better classification.

3 Experiments

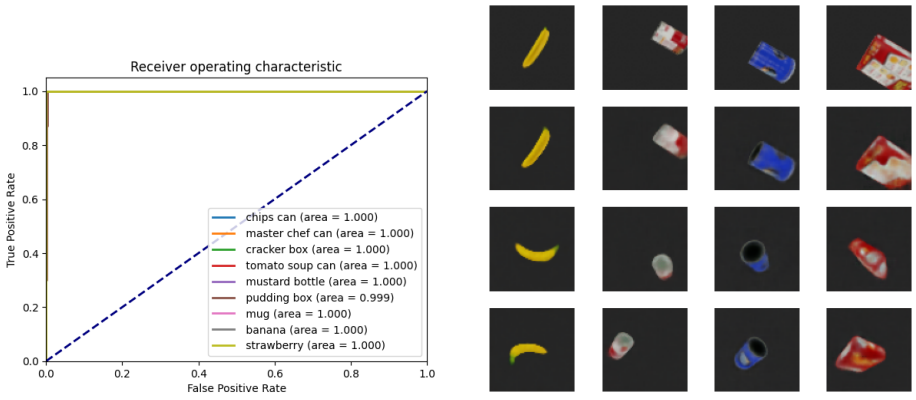
We evaluate our model for an agent in a 3D environment, where 3D models of objects from the YCB dataset [1] are rendered from a certain camera viewpoint. The agent actions are then defined as relative transforms (i.e. rotation and translation), moving the camera viewpoint. This setup closely mimicks a robot manipulator with an in-hand camera, but without kinematic constraints [26].

We create a dataset using 3D meshes of objects from the YCB dataset [1]. For each of 9 "known" objects, 14000 viewpoints and their corresponding view, for which the object is centered in view, are generated as a train set. During training, pairs of two views are randomly selected, for which the action is defined as the relative transform between these two viewpoints.

We first validate that the CCN ensemble is able to learn pose and identity representations unsupervisedly by minimizing free energy. Next, we show how the expected free energy allows to agent to infer actions that can bring the agent to a preferred pose relative to an object on the one hand, and resolve ambiguity for inferring an object identity on the other hand.

3.1 The "what" stream: object recognition

First, we evaluate the performance of each individual CCN "what" binary classifier. The ROC curves are shown in Figure 2a where each CCN is tested on a dataset with 3000 novel views for each of the 9 known objects, and 3000 views from 5 objects, it has never seen during training. For all objects we achieve near-perfect ROC curves, which can be attributed to the fact that each CCN



(a) (b)

Fig. 2: (a) The ROC curve for individual CCNs. Negative samples are observations from the test set: padlock, power drill, knife, orange and tuna fish can. (b) The ground truth (top row), the reconstruction (second row), and imagined transformed observations (other rows) are shown for multiple YCB objects [1].

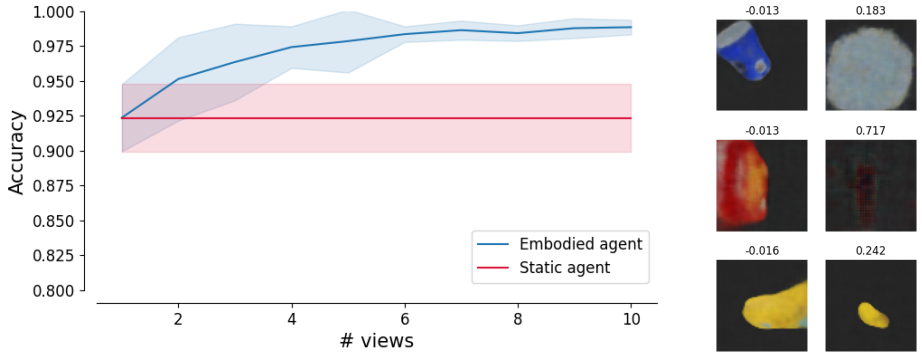
can focus on particular features that distinguish a particular object from the others. Investigating the impact on the ROC performance when using real-world observations instead of 3D renders of predefined object models would be an interesting avenue for future work.

3.2 The "where" stream: implicit pose estimation

Crucially, our CCNs not only learn a classification output, but also an implicit representation of the 3D structure of the object at hand. As discussed in Section 2, this is encoded in a latent code $\mathbf{v}_t$, from which the model can reconstruct the given viewpoint using the decoder, or imagine other viewpoints after a relative transform using the transition model. This is illustrated in Figure 2b, where the first row shows ground truth object observations, the second row shows the reconstruction after encoding, and the third and fourth row show imagined other viewpoints.

We can now use the CCN to infer the actions that will yield some "preferred" observation, by minimizing the expected free energy in Equation 3. This is useful for example to instruct a robotic manipulator to a certain grasp point for an object. As computing G for every action is intractable, we sample 1000 random relative transforms for which G is calculated. A transform is sampled by first sampling a target viewpoint in 3D space uniformly in the workspace. The orientation is then determined so that the camera looks at the center of gravity of the object. The relative transform can then be computed between both current and target sampled poses. The identity transform is always provided as an option, allowing the agent to stay at its current pose when no better option is found. This results in the agent finding the estimated pose. Figure 3 shows qualitative trajectories for estimating the correct pose from both a mug and a pudding box. On average, the pose estimation process converges after 3 steps, and the resulting final pose lies around 1 mm (average of 1.4 mm) and 5 degrees (average of 4.7 degree) in distance and angle compared to the ground truth. We provide a more detailed table in Appendix B.

Fig. 3: The agent is provided with a preferred observation (first column left), and an intial observation (second column). The agent infers the relative transforms to reach the preferred observation.



(a) (b)

Fig. 4: (a) Performance of the CCN ensemble in a static (red) and an embodied agent, driven through active inference (blue). The agent is provided with 20 situations for each of the 9 known objects and 5 never-before seen test objects. (b) Imagined views for actions that result in the largest and smallest expected free energy G.

3.3 Embodied agents for improved classification

Whereas previously we evaluated the binary classfication performance of individual CCNs, we now evaluate the performance of the CCN ensemble as an $n+1$-way classifier, with n object classes and one "other" class.

We evaluate an embodied agent that can query extra viewpoints to improve its classification. In this case, the agent again infers actions that minimize the expected free energy G, effectively maximizing information gain on the object identity. In this case, CCN votes are aggregated in the concentration parameters of a Dirichlet distribution, as described in Section 2. We also add a fixed 0.5 value vote for the "other" object category which accounts for evindence for the "other" class when none of the CCNs "fire". We compare this agent with a static agent that only has a single view. In this case, only a single vote is used for the parameters of the Dirichlet distribution.

To evaluate the performance, we randomly sample 20 views for each of 14 object classes (9 known object classes and 5 never seen before), and evaluate both the static and embodied classificaton accuracy. The results are shown in Figure 4a. Whereas the static agent achieves an overall accuracy of 92.5%, the embodied agent consistently improves in accuracy, reaching 98%, as more viewpoints are queried, in line with [8]. The error bounds are computed over 5 different random seeds and represent the 95% HDI. Figure 4b shows the imagined observations for the largest and smallest expected free energy G. Figure 4b shows imagined views with highest or lowest G, and illustrates that the active inference agent prefers observations where the object is clearly in view from a more close up view, rather than more ambiguous viewpoints.

4 Related work

Deep learning has been widely used for static image classification [15, 10]. However, recent work also focused on active vision. In [26] a generative model learning representations of a whole 3D scene was used for an active inference agent, whereas in [2] an explicit what and where stream were modeled for classifying MNIST digits.

Recently a lot of progress has been made in methods that learn the 3D geometry of objects. The geometry can either be learned implicitly using Neural Radiance Fields (NeRF) [16] or Generative Query Networks (GQN) [3] or explicitly using Scene Representation Networks (SNR) [24]. However these approaches either have to generalize to a large variety of objects, which results in an involved training process requiring a lot of data, or they optimize for a single observation, limiting the flexibility.

Continual learning methods are able to use experience gathered during deployment of a system to improve the system over time. Typical approaches involve an ensemble of classifiers, that operate on a subset of the inputs, either by splitting the train data in specific subsets to train a mixture of experts [12], or by identifying clusters in a shared latent space and training separate classifiers for separate clusters [23].

The use of information gain has also been used as an exploration strategy outside of the active inference community, in which it substantially improves exploratory performance on a number of Atari tasks [17].

While most approaches tackle these problems separately, we propose a biologically inspired method that learns object-centric representations in an unsupervised manner for both the object identity and its geometric properties.

5 Discussion

We believe that this is a first step towards manipulation of three dimensional objects, and plan to extend this work to a real-world robot setup. In this case, the robotic agent needs to make inferences on the object pose and identity that is present in the workspace. In case the object is identified, the agent is attracted to preferred observations, e.g. for grasping or manipulating the object. In case a different, novel object class is identified, a new CCN is instantiated and trained on these novel object views. In this case, we could infer the viewpoints that have a high information gain on model parameters in an active learning setting, which can also be written as an expected free energy term [5]. We can further extend the generative model to also take into account multiple objects in the scene and modelling inter-object relations and geometry.

A limitation of our current setup is that it can only deal with a single object in the center of the view. As multi-object scenes are ubiquitous in the real world, this is a natural direction for future work. We propose a solution in which the agent can divide its spatial attention on the observations by looking at the CCN activations at different patches on the observation. Once an object and its

relative reference frame is found, these can be linked using a global, ego-centric reference frame of the agent [18]. This way, a hierarchical generative model of the whole workspace, composed of different object is constructed. These latent parameters can then be propagated over time through a predictive model, and can in that way deal with occlusions.

In principle, one could instantiate a hierarchy of CCNs, where higher level CCNs process the output of lower level CCNs, effectively modeling part-whole relationships. This is similar to Capsule Networks [21] and GLOM [11], and corresponds better with the 1000 brains theory [9]. However, given the limited scalability of state of the art implementations of such hierarchical approaches [21], we adopted CCNs that operate on the level that is most important for a robot operating in a workspace, i.e. the discrete object level.

We found that failure cases exist when CCNs incorrectly "fire" for unseen objects. This confusion occurs for some objects yielding a non-perfect classification score. We could further improve the system by also taking into account how good novel observations match our predictions in the past.

6 Conclusion

In this paper we showed a novel approach to modelling 3D object properties, drawing inspiration from current development in the Neuroscientific domain. We proposed to model a separate what- and where stream for each individual object and are able to use these models for object identification as well as implicit object pose estimation. We show that through embodiment, these models can aggregate information and increase classification performance. Additionally, we show that by following the free energy formulation, these module networks can be used for implicit pose estimation of the objects.

Acknowledgments This research received funding from the Flemish Government (AI Research Program). Ozan Çatal was funded by a Ph.D. grant of the Flanders Research Foundation (FWO). Part of this work has been supported by Flanders Innovation & Entrepreneurship, by way of grant agreement HBC.2020.2347.

References

1. Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015.
2. Emmanuel Daucé and Laurent U Perrinet. Visual search as active inference. In *IWAI 2020*, 2020.
3. S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals,

Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, June 2018.
4. George Ettlinger. "object vision" and "spatial vision": The neuropsychological evidence for the distinction. *Cortex*, 26(3):319–341, 1990.
5. Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016.
6. Justin Gilmer, Ryan P. Adams, Ian J. Goodfellow, David G. Andersen, and George E. Dahl. Motivating the rules of the game for adversarial example research. *CoRR*, abs/1807.06732, 2018.
7. Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, December 2020.
8. Jeff Hawkins, Subutai Ahmad, and Yuwei Cui. A theory of how columns in the neocortex enable learning the structure of the world. *Frontiers in Neural Circuits*, 11:81, 2017.
9. Jeff Hawkins, Marcus Lewis, Mirko Klukas, Scott Purdy, and Subutai Ahmad. A framework for intelligence and cortical function based on grid cells in the neocortex. *Frontiers in Neural Circuits*, 12:121, 2019.
10. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.
11. Geoffrey E. Hinton. How to represent part-whole hierarchies in a neural network. *CoRR*, abs/2102.12627, 2021.
12. Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.
13. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
14. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
15. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.
16. Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.
17. Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning, 2019.
18. Thomas Parr, Noor Sajid, Lancelot Da Costa, M. Berk Mirza, and Karl J. Friston. Generative models for active vision. *Frontiers in Neurorobotics*, 15:34, 2021.
19. Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014.
20. Danilo Jimenez Rezende and Fabio Viola. Taming vaes, 2018.
21. Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *CoRR*, abs/1710.09829, 2017.
22. Adam Safron. The radically embodied conscious cybernetic bayesian brain: From free energy to free will and back again. *Entropy*, 23(6), 2021.

23. Murray Shanahan, Christos Kaplanis, and Jovana Mitrovic. Encoders and ensembles for task-free continual learning. *CoRR*, abs/2105.13327, 2021.
24. Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019.
25. Ryan Smith, Karl Friston, and Christopher Whyte. A step-by-step tutorial on active inference and its application to empirical data, Jan 2021.
26. Toon Van de Maele, Tim Verbelen, Ozan Çatal, Cedric De Boom, and Bart Dhoedt. *Frontiers in Neurorobotics*, 15:14, 2021.

Appendix A Neural Network Architecture and Training Details

The neural network is based on a variational autoencoder [14, 19] consisting of an encoder and a decoder. The encoder ϕ_θ uses a convolutional pipeline to map a high dimensional input image (64x64x3) into a low dimensional latent distribution. We parameterize this distribution as a Bernouilli distribution representing the identity of the object and the camera viewpoint as a Multivariate Normal distribution with diagonal covariance matrix of 8 latent dimensions. The decoder ψ_θ then takes a sample from the viewpoint and is able to reconstruct the observation through a convolutional pipeline using transposed convolutions. In addition to a traditional variational autoencoder, we have a transition model χ_θ that transforms a sample from the viewpoint distribution to a novel latent distribution, provided with an action. This action is a 7D vector representing the translation as coordinates in and rotation in quaternion representation. The model architecture for encoder, decoder and transition models are shown in Table 1, Table 2 and Table 3, respectively.

The model is optimized end-to-end through the minimization of Free Energy as described in Equation 2. The expectations over the different terms are approximated through stochastic gradient descent using the Adam optimizer [13]. As minimization of negative log likelihood over reconstruction is equivalent to minimization of the Mean Squared Error, this is used in practice. Similarly, the negative log likelihood over the identity is implemented as a binary cross-entropy term. We choose the prior belief over $\mathbf{v}$ to be an isotropic Gaussian with variance 1. The individual terms of the loss function are constrained and weighted using Lagrangian multipliers [20]. We consider only a single timestep during the optimization process. In practice this boils down to:

$$\begin{aligned} L_{FE} =& \lambda_1 \cdot L_{BCE}(\hat{i}, i) + \lambda_2 \cdot L_{MSE}(\psi_\theta(\hat{\mathbf{v}}_{t+1}), \mathbf{o}_{t+1}) \\ &+ D_{KL}[\ \underbrace{\chi_\theta(\mathbf{v}_t, \mathbf{a}_t)}_{q(\mathbf{v}_{t+1}|\mathbf{v}_t, \mathbf{a}_t, \mathbf{i})}\ ||\ \underbrace{\phi_\theta(\hat{\mathbf{o}})}_{p(\mathbf{v}_{t+1}|\mathbf{i}, \mathbf{o}_t)}\ ] \end{aligned} \tag{4}$$

where $\hat{i}$ is the prediction $\phi_\theta(\mathbf{o}_t)$ of the what-stream for the encoder, $\hat{\mathbf{v}}_{\mathbf{t+1}}$ is a sample from the predicted transitioned distribution $\chi_\theta(\mathbf{v}_t, \mathbf{a}_t)$ and $\hat{\mathbf{o}}_{t+1}$ is the expected observation from viewpoint $\hat{\mathbf{v}}_{\mathbf{t+1}}$, decoded through $\psi_\theta(\mathbf{v}_{t+1})$. The λ_i variables represent the Lagrangian multipliers used in the optimization process.

During training, pairs of observations $\mathbf{o}_t$ and $\mathbf{o}_{t+1}$ and corresponding action $\mathbf{a}_t$ are required. To maximize data efficiency, the equation is also evaluated for zero-actions using only a single observation, and reconstructing this directly without transition model.

Table 1: Neural network architecture for the image encoder. All strides are applied with a factor 2. The input image has a shape of 3x64x64. The output of the convolutional pipeline is used for three different heads. The first predicts the mean of the distribution μ, the second head predicts the natural logarithm of the variance σ^2, for stability reasons and finally the third head predicts the classification output score c as a value between zero and one after activation through the sigmoid activation function.

Output label	**Layer**	**Kernel size**	**# Filters**
	Strided Conv2D	4	8
	LeakyReLU		
	Strided Conv2D	4	16
	LeakyReLU		
	Strided Conv 2D	4	32
	LeakyReLU		
	Strided Conv2D	4	64
	LeakyReLU		
h	Reshape to 128		
μ	Linear (input: h)		8
$\ln \sigma^2$	Linear (input: h)		8
c	Linear + Sigmoid (input: h)		1

Table 2: Neural network architecture for the image decoder. The input of this model is a sample drawn from the latent distribution, either straight from the encoder, or transitioned through the transition model. All transpose layers use a stride of two. The final layer of the model is a regular convolution with stride 1 and kernel size 1, after which a sigmoid activation is applied to map the outputs in the correct image range.

Layer	**Kernel size**	**# Filters**
Linear		128
Reshape to 128x1x1		
ConvTranspose2D	5	64
LeakyReLU		
ConvTranspose2D	5	64
LeakyReLU		
ConvTranspose2D	6	32
LeakyReLU		
ConvTranspose2D	6	16
LeakyReLU		
Conv2D	1	3
Sigmoid		

Table 3: Neural network architecture for the transition model. The input from this model is an 8 dimensional latent code, concatenated with the 7-dimensional representation of the relative transform (position coordinates and orientation in quaternion representation). For stability reasons, the log-variance is predicted rather than the variance directly.

Output label	**Layer**	**# Filters**
	Linear	128
	LeakyReLU	
	Linear	256
	LeakyReLU	
	Linear	256
	LeakyReLU	
μ	Linear	8
$\ln \sigma^2$	Linear	8

Appendix B Additional experimental details

In Table 4, the computed angular and translational distances for the 9 evaluated objects are shown. Figure 5 shows a sequence of imaginations for all 9 objects, the top row represents the ground truth input, the second row the reconstruction and the subsequent rows are imagined observations along a trajectory.

Table 4: The mean distance error in meters and angle error in radians for different objects of the YCB dataset [1] in our simulated environment. For each object 20 arbitrary target poses were generated over which the mean values are computed.

Object	**Distance error (m)**	**Angle error (rad)**
chips can	0.00328 ± 0.00824	0.15997 ± 0.21259
master chef can	0.00036 ± 0.00034	0.06246 ± 0.03844
cracker box	0.00028 ± 0.00023	0.04659 ± 0.02674
tomato soup can	0.00073 ± 0.00104	0.08653 ± 0.07021
mustard bottle	0.00070 ± 0.00072	0.06351 ± 0.03818
mug	0.00083 ± 0.00128	0.09098 ± 0.10232
pudding box	0.00051 ± 0.00052	0.06190 ± 0.03843
banana	0.00055 ± 0.00042	0.07482 ± 0.03592
strawberry	0.00573 ± 0.01181	0.16699 ± 0.15705

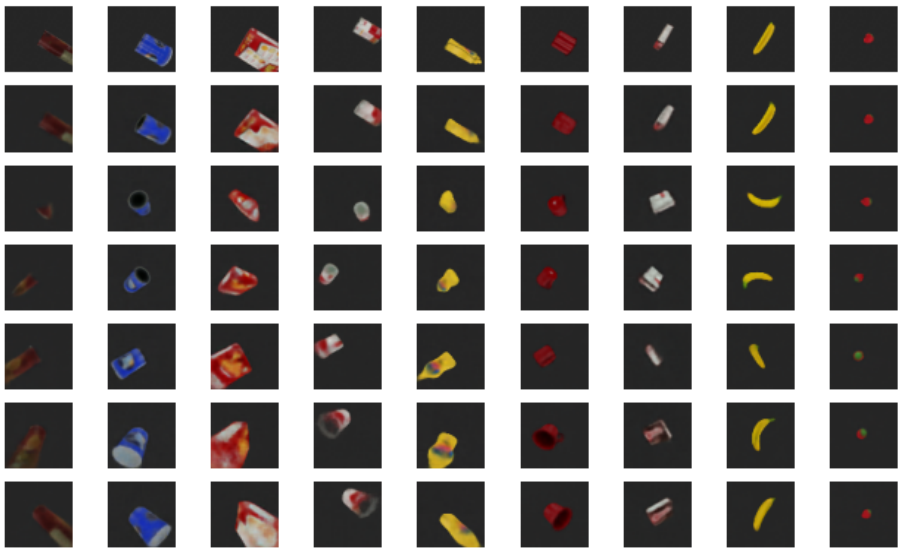


Fig. 5: The top row represents the ground truth observation that was provided as input to the model. The second row shows a direct reconstruction when no action is applied to the transition model. All subsequent rows show imagined observations along a trajectory.

Rule learning through active inductive inference

Tore Erdmann[1] and Christoph Mathys[2,3]

[1] SISSA, Via Bonomea 265, 34356 Trieste Italy
terdmann@sissa.it
[2] Interacting Minds Center, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus, Denmark
[3] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

Abstract. We propose a grammar-based approach to active inference based on hypothesis-driven rule learning where new hypotheses are generated on the fly. This contrasts with traditional approaches based on fixed hypothesis spaces and Bayesian model reduction. We apply these two contrasting approaches to an established active inference task and show that grammar-based agents' performance benefits from the explicit rule representation underpinning hypothesis generation. Our proposal is a synthesis of the active inference framework with language-of-thought models, which paves the way for computational-level descriptions of false inference based on an aberrant hypothesis-generating process.

Keywords: active inference · rule induction · context free grammars · structure learning · sampling-based inference · reasoning

1 Introduction

Structure learning is a fundamental problem for an active inference agent. Logically structured concepts can be found in domains such as mathematics, social systems or causal processes [13]. The likelihood mapping of a POMDP with discrete state space can be represented as a matrix with elements indicating the likelihood of an observation given a state. Current approaches for learning this mapping rely on separately estimating the individual elements of the matrix [4,12]. Here, we propose an approach for structure learning that uses a prior based on context-free grammars (CFG; [2]), which were invented in linguistics to describe the structure of sentences in natural language and are used to define programming languages in computer science. From such a grammar, the agent can, through recursive composition and substitution of terms, generate an infinite number of expressions, which represent the underlying structure of (parts of) its environment. As a proof of concept, we will illustrate our approach by applying it to a rule learning problem inspired by the task in [4].

This approach has previously been used in cognitive science, psychology [13,6,14] and, in particular, in "language of thought" models [10]. Previous work

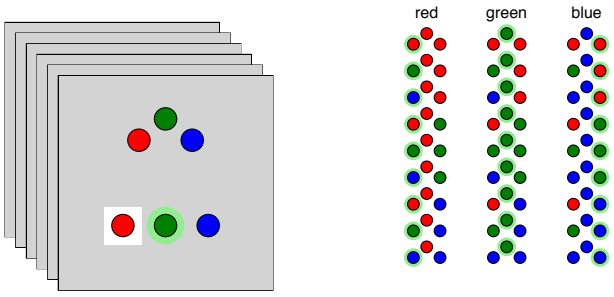


Fig. 1: Left: shows the display during a trial. The agent sees context variables (the three circles in the upper half), makes a response (indicated by the white box around the red circle) and, having made a choice, the correct choice (highlighted in green) . Right: The possible contexts arranged according to the value of the middle circle, which implies where to look for the correct choice (highlighted in green). The correct response is equal to the color of the circle on the left, in the center or on the right when the color of the central circle is "red", "green" or "blue", respectively.

has shown that these models can account for various features of human concept learning ([5]). Furthermore, this approach has been used to explain surprise signals in the striatum [1].

We will work through a simplified version of the task of [4]. For ease of presentation and to place our focus on structure learning, we remove state uncertainty and all intra-trial actions except for the final choices. All remaining uncertainty is thus about the hidden rule. However, our proposal can be straightforwardly applied to the case including state uncertainty and observations corrupted by noise. In this task, see Fig. 1, the agent has to infer a rule, that is a deterministic mapping from three context variables to the correct choice.

2 Active inference

Solving this task consists of finding a policy $p(a_t|c_t, \theta)$, that gives the probability of a choice $a_t \in \{1, 2, 3\}$ given the context variables c_t and some parameters θ. The generative model the agent holds of the task is

$$c_t^{(j)} \sim U(\{1, 2, 3\}),\ j = 1, 2, 3 \tag{1}$$

$$c_t = (c_t^{(1)} c_t^{(2)}, c_t^{(3)}) \tag{2}$$

$$o_t \sim f(c_t, \cdot) \tag{3}$$

$$p(r_t = 1|a_t, o_t) = \exp(\ell(a_t, o_t)) \tag{4}$$

$$\ell(a_t, o_t) = \begin{cases} 0 & \text{if } a_t = o_t \\ -4 & \text{else} \end{cases} \tag{5}$$

where $f(c, o)$ is a function representing the hidden rule. That is, it returns the probability of observing the outcome $o \in \{1, 2, 3\}$ in context c. The prior about reward observations $p(r_t|a_t, o_t)$ represents an optimistic bias, so that the agent's beliefs are biased by desirable states and not the actual task dynamics, which is $r_t = \mathbb{I}(o_t = a_t)$. This model implies a distribution over trial sequences, which we denote $\tau = (c_{1:T}, a_{1:T}, o_{1:T}, r_{1:T})$, that factorizes as

$$p(\tau|f) = \prod_{t=1}^{T} p(r_t|a_t, o_t)p(o_t|c_t, f)p(a_t)p(c_t). \tag{6}$$

Given the biased prior over rewards we obtain the following posterior over actions when conditioning on $r_t = 1,\ \forall t = 1, \ldots, T$ and summing out o_t, which is unknown at the time of the action,

$$p(c_{1:T}, a_{1:T}, o_{1:T}|r_{1:t} = 1, f) \propto \prod_{t=1}^{T} p(r_t = 1|a_t, o) \tag{7}$$

In keeping with the active inference framework, the expected log model evidence is minimized by computation of the posterior over action, which can be done at each trial t by choosing

$$p(a_t|r_t = 1, c_{1:t}, o_{1:t-1}) = \sigma(-G_{a_t}) \tag{8}$$

$$G_{a_t} = \sum_{o} l(a_t, o) \cdot \mathbb{E}_{p(f|c_{1:t-1}, o_{1:t-1})}\Big[p(O_t = o|c_t, f)\Big] \tag{9}$$

For the implementation, this means we need to be able to evaluate the agent's posterior predictive about the belief about the outcome o_t. The above constructions leads to the maximization of the following objective (see [7])

$$D_{KL}(p^*(\tau)||p(\tau)) = \mathbb{E}_{\tau \sim p^*(\tau)}\Big[\log p(\tau) - \log p^*(\tau)\Big], \tag{10}$$

which is the Kullback-Leibler divergence of the agent's beliefs about its future states and a desired distribution over these p^*, and which is equivalent to the free energy of the expected future, which is a lower bound on the expected log model evidence [8].

2.1 Evidence accumulating agent

A straightforward solution for learning the rule is available if we represent it as a stochastic vector consisting of independent Dirichlet variables, $f(c, o) = \theta_{c,o}$, with $\theta_{j,\cdot} \sim Dir(\alpha_0),\ j = 1, \ldots, 27$, for which the posterior can be computed by accumulation of concentration parameters:

$$p(\theta_{j,o}|c_{1:t}, o_{1:t}) = Dir(n_{c,o} + \alpha_0) \tag{11}$$

where $n_{c,o}$ is the number times (up until time t) the agent has observed outcome o for context c. If we define a matrix α with entries $\alpha_{c,o} = n_{c,o} + \alpha_0$, the expectation

in eq. 8 is a that of a categorical-Dirichlet distribution and the action is chosen via

$$G_{a_t} = \sum_o \ell(a_t, o)) \cdot \frac{\alpha_{c_t,o}}{\sum_j \alpha_{c_t,j}}. \tag{12}$$

2.2 Bayesian model reduction

If the agent knows that there must be a deterministic rule, it can quickly recognize the rule by comparing the evidence for each potential model in a set of hypothetical models and accept a model if its evidence exceeds a certain threshold.

The model space can be considered the set of deterministic, one-to-one mappings from each color to each response (of which there 6) which are combined with the 6 possible mappings between the central color and which location the color-to-response mapping should be applied to (see [4]). There are thus 36 hypotheses, for which the evidence is computed on each trial. This allows us to represent the priors through sets of prior concentration parameters as derived in [4]. A condition for this agent is that the space of hypotheses is specified for the agent beforehand, which is a strong assumption in general. We will now introduce a way to model acquisition of new models. This has the advantage of being based on weaker assumptions about (and a different conception of) prior knowledge.

3 Grammar-based rule induction

Here, we describe how rule learning can be supported through a structured prior over an auxiliary space of symbolic rule expressions. Each such rule expression is defined by a syntax tree, consisting of logical connectives (and, or), and references to the observations in a trial. The "leaf nodes" of the tree are predicates of some part of the observation c_t, for example $color(c_t^{(1)}) = red$, which is either true or false (see appendix for an example). An agent can learn a rule expression that accurately predicts the outcome of the unknown rule f by searching the space of rule expressions for hypotheses which are then evaluated against the available evidence. Hypotheses are represented by expressions that can be generated by iterating the following set of re-write (or production) rules:

$$\begin{aligned}
\text{(Start)}\ & S \rightarrow f(c,o) \iff (D) \\
\text{(Disjunction)}\ & D \rightarrow C \vee D \mid P \mid false \\
\text{(Conjunction)}\ & C \rightarrow P \wedge C \mid P \mid true \\
\text{(Predicate)}\ & P \rightarrow color(Loc) = Col \\
\text{(Location)}\ & Loc \rightarrow c_1 \mid c_2 \mid c_3 \\
\text{(Color)}\ & Col \rightarrow \text{“}red\text{”} \mid \text{“}green\text{”} \mid \text{“}blue\text{”}
\end{aligned}$$

These rules indicate how symbols on the left hand side of the $\rightarrow$ can be replaced by one of the options on the right hand side (options are separated by |). From this grammar, given certain production probabilities (which give the probability of each possible production for each line in the grammar; can be assumed uniform), we can generate rule expressions (we refer the interested reader to Wikipedia, for examples, or [11] for a comprehensive treatment). Note that we omit the trial index t in the formulas (since the rules only refers to variables in the current trial) and instead use the subscript to denote the location (1, 2 or 3) of the context variable.

Each generated expression describes some arrangement of context observations. Say, we wanted to describe the rule for when the correct color is red (as given in the caption of 1). This can be expressed as $color(c_2) = \text{“}red\text{”} \wedge color(c_1) = \text{“}red\text{”}) \vee (color(c_2) = \text{“}blue\text{”} \wedge color(c_3) = \text{“}red\text{”})$, which can be generated through step-wise replacement of the above rules. The prior probability of a formula (i.e. a sequence of substitutions from the grammar) is equal to the product of the probabilities of the individual substitutions. This prior naturally places higher probability on shorter and less complex expressions since they include fewer terms in the product.

For the rule learning task described above, we want to model the contexts that correspond to the three outcomes (and actions), so we will make the procedure to be learned a function of both the observed context c and the outcome o, changing the rule in the topmost line above to be a context-sensitive expression of the form

$$S \rightarrow f(c, o) \iff ((o = \text{“}red\text{”}) \wedge D) \vee ((o = \text{“}green\text{”}) \wedge D) \vee ((o = \text{“}blue\text{”}) \wedge D),$$

wherein the D terms will come to represent the parts of the rule that imply the corresponding outcome. We can then evaluate expressions with regard to each possible outcome to determine if the context c matches the outcome o. Starting from the above expression and generating sub-expressions according to the above grammar, we can represent the true hidden rule described in Fig. 1 as follows:

$$\begin{aligned} f(c, o) \iff & ((o = \text{“}red\text{”}) \wedge \\ & ((color(c_2) = \text{”}red\text{”} \wedge color(c_1) = \text{”}red\text{”}) \vee \\ & (color(c_2) = \text{”}blue\text{”} \wedge color(c_3) = \text{”}red\text{”}))) \\ & \vee ((o = \text{“}green\text{”}) \wedge \\ & ((color(c_2) = \text{”}green\text{”}) \vee (color(c_2) = \text{”}red\text{”} \wedge color(c_1) = \text{”}green\text{”}) \vee \\ & (color(c_2) = \text{”}blue\text{”} \wedge color(c_3) = \text{”}green\text{”}))) \\ & \vee ((o = \text{“}blue\text{”}) \wedge \\ & ((color(c_2) = \text{”}red\text{”} \wedge color(c_1) = \text{”}blue\text{”}) \vee \\ & (color(c_2) = \text{”}blue\text{”} \wedge color(c_3) = \text{”}blue\text{”}))) \end{aligned}$$

However, we can represent this rule more succinctly by adding more abstract terms to the grammar. For example, by adding two new production rules to the grammar above:

$$P \to color(Loc) = COL \;\mid\; o = color(Loc)$$
$$Loc \to c_1 \;\mid\; c_2 \;\mid\; c_3 \;\mid\; c_{Loc}$$

The last production will lead to a "subsetting", such as c_{c_2}, which means that the value of c_2 indexes the context variables (with the colors mapped to the numbers $\{1, 2, 3\}$). The expression $o = color(Loc)$ evaluates to true if the outcome matches the variable Loc. With these additions, we can now represent the true rule as a much shorter expression

$$f(c, o) \iff (o = color(c_{c_2})). \tag{13}$$

This shorter representation of the rule helps the agent to discover it much more quickly. This is because shorter rules have higher prior probabilities of being produced.

The above rule expression defines a function that evaluates to `true` if the action a is correct given the observation o and `false` otherwise. The likelihood of this expression is given by its match with the observed data, that is, the number of examples for which the rule f evaluates to true,

$$p(f|o_{1:t}, a_{1:t}, c_{1:t}) \propto \bigwedge_{c,o} f(c, o) \tag{14}$$

or, if assuming that some observations might be outliers to the rule, we have

$$p(f|o_{1:t}, a_{1:t}, c_{1:t}) \propto e^{-\gamma Q(f)} \tag{15}$$

where $Q(f) = |\{(c, o) \in (c_{1:t}, o_{1:t}) : f(c, o) = false\}|$ (the count of examples for which the rule expression evaluates to false) and γ is a parameter denoting the probability that a given example is an outlier. Here, the probabilities need not be normalized, since any normalization constants cancel in the MCMC acceptance probability. The truth value of the procedure $f(a, o)$ follows from the evaluation approach in mathematical logic [3] and is defined recursively:

1. $f(a, o)$ is a node.
2. If a node is a predicate, it can be evaluated directly
3. If it is a logical connective then it is evaluated by first evaluating the sub-expressions separately and then applying the logical function to the result. For example, $a \wedge b$ is true only if both sub-expressions a and b are true.

In our implementation, we represent the agent's belief about the correct rule expression as a set of samples that are approximately distributed according to the posterior distribution implied by the above likelihood and prior. This posterior is updated on each trial by running a Markov Chain Monte Carlo (MCMC) chain for a fixed number of iterations. The set of expressions that was visited during

the walk is taken to represent the posterior belief. This construction leads to the posterior predictive distribution, given a set H_t of hypotheses. Formally, if we denote the chain representing the belief update in trial t by $H^{(t)} = \left(h_1^{(t)}, \ldots, h_n^{(t)}\right)$, we can evaluate the posterior expectation in action selection in eq. 8 approximately as follows

$$p(a_t = a | r_t = 1, c_{1:t}, o_{1:t-1}) = \sigma(-G_{a_t}) \tag{16}$$

$$G_{a_t} = \sum_o l(a_t, o) \cdot \mathbb{E}_{p(f|c_{1:t-1}, o_{1:t-1})}\Big[p(O_t = o | c_t, f)\Big] \tag{17}$$

$$\approx \sum_o \ell(a_t, o)) \cdot \frac{\sum_i f_{h_i^{(t)}}(c_t, a)}{\sum_{j \in \{1,2,3\}} \sum_i f_{h_i^{(t)}}(c_t, j)} \tag{18}$$

which can be seen as a model average of all hypotheses that were visited by the Markov chain during the computation of the posterior.

The iterations of the MCMC procedure propose changes to the expression by randomly selecting a sub-expression and replacing it with a newly generated sub-expression. The Metropolis-Hastings acceptance probability for a proposal balances the probability of the proposal and the reverse proposal, the prior probabilities and the likelihood (see eq. 14) of the current and proposed expressions (tree-substitution MCMC; see [5] for details). The belief update can thus be performed by running n MCMC iterations, starting from the current state of the chain. For the current task, once the true rule has been found, proposal for moves away from it will have very low probability. In general, when the rule cannot be known with certainty, the chain will move between alternatives and thereby lead to a representation of the remaining uncertainty in the posterior belief about the rule.

4 Experiments

We simulated learning in four agents who completed 20 trial sequences each. These sequences contained 27 trials and were generated by randomly shuffling the 27 unique combinations of context variables. The four agents differed in substantial ways and could be characterized as concentration parameter accumulating agents (Agents 1 and 2, described in section 2.1) with (Agent 2) and without (Agent 1) model-selection (by Bayesian model reduction, sec. 2.2) after each trial; and the grammar-based agents (Agents 3 and 4) with the simple grammar described in 3 (Agent 3) and an extended grammar described below (Agent 4).

A comparison of the performance of the different agents are shown in Figure 2, where the average proportion of correct responses is shown over trials. As can be seen, the grammar-based agents show higher proportions of correct responses already during early trials. This is due to the nature of the rule expressions, which can be extrapolated from rapidly.

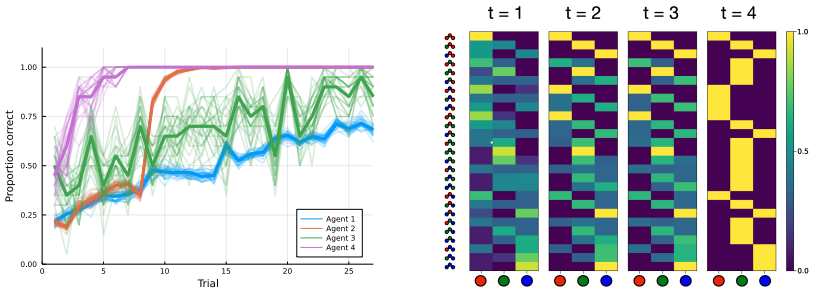


Fig. 2: Left: Proportion of correct choices (averaged over simulations) for the four agents with uncertainty indicated via bootstrapped estimates (thin lines). Right: Belief of (purple) grammar-based agent for the first 4 examples of a particular trial sequence. Each heatmap shows the probability of an action (x-axis) to be correct in a given context (y-axis).

The best-performing agent (purple in Fig 2) is Agent 4 with the extended grammar, that has two additional production rules contained in its grammar (see sec. 3). Figure 2 (right) shows the Agent 4's belief about the rule during the early trials of a particular trial sequence. The examples presented to the agent were ((●, ●, ●), (●, ●, ●), (●, ●, ●), (●, ●, ●)), for which the correct responses were (●, ●, ●, ●). We can inspect the set of hypotheses held by the agent. At $t = 3$, the hypotheses with the highest weights (about $n/3$ occurrences) are:

1. $a = color(o_1)$, "answer equal to the left circle"
2. $a = color(o_{o_2})$ "answer equal to the color at location indicated by o_2"
3. $a = color(o_{o_3})$ "answer equal to the color at location indicated by o_3"

The agent cannot tell between these explanations until observing the outcome in the 4th trial, when the predictions of hypotheses 1 and 3 are disproved and the agent correctly infers the rule.

These results show how learning speed relates to underlying assumptions. As opposed to Agents 1 and 2, who need to be equipped with a fixed hypothesis set, the grammar-based agents can learn arbitrary rules, including such for which maintaining a fixed hypothesis set would be infeasible, as long as they can be represented within the language spanned by their grammar. For example, if we had just told the agent: "In this game, there is a deterministic mapping between the three colored circles and the correct response", the hypothesis space would have to cover a space of mappings containing 2^{81} elements. Comparing each of these candidates at the end of a trial would be infeasible (at a rate of 10^9 evaluations per second (1 evaluation per nanosecond), it would take about 77 million years to evaluate all candidates). The code for all experiments reported here is available at https://github.com/ilabcode/IWAI2021.

5 Discussion

We have shown a novel way to perform structure learning in active inference agents. In particular, we demonstrate how an agent can use grammar-based structure learning to develop a model in a bottom-up fashion. This is different from the traditional approach of Bayesian model reduction, which can be considered a top-down approach. The assumption of a grammar that spans a hypothesis space is weaker and hence more generalizable than pre-defining a finite set of hypotheses. Other ways of searching for rule expressions are possible, such as genetic algorithms, but these do not represent uncertainty and are therefore not well suited as a basis for adaptive prediction and decision-making.

Our results showed differences between the two grammar-based agents that were apparent in the speed by which they learn the rule. For the task presented here, both agents converge to the same behavior, but their underlying rule representations are different. This highlights how higher-order inferences can depend on the base of concepts and abstractions they are built upon. In terms of the behavior, the agents will look the same, however, their representational vocabulary differs and so they will find separate explanations for the rule (which do describe the same contingencies), which also have different complexity (as clearly visibly in the number of terms). Given a way to update their own grammars through experience, two agents starting with different grammars but in similar environments might develop a similar conceptual toolbox. One way to enable this would be to add special "lambda expression" terms to the grammar. Such an encoding of the lambda calculus within the hypothesis language leads to the ability to define new terms and apply or re-combine them (see [9]).

An interesting aspect of your hypothesis-generating grammar-based approach is the ways in which the assumptions underlying the generation of hypotheses of can influence what the agent finally takes to be the most promising course of action. This can become a useful tool for understanding aberrations in world modeling such as those apparent in psychiatric illnesses, which might have to do with a deficient hypothesis-generating process. For example, hypotheses generated from a grammar that is poorly attuned to a domain can seem bizarre to outside observers. Such misattunement may be the result of aberrant learning processes that update the production probabilities of a grammar, or the addition or removal of terms.

The agent described in [4] did not include model-selection considerations in its actions since they were outside of its generative model (and, in any case, the actions in the task were uninformative in that regard). By contrast, with a grammar-based approach, the structure is part of the agent's prior. Therefore its actions can subserve the testing of freshly generated hypotheses about the hidden structure of a task, which corresponds to active learning. Crucially, this could be made relevant in a version of the rule learning task where the agent can choose its next set of context variables. This would require planning, where the agent finds the optimal plan for testing its currently most promising hypotheses — an interesting avenue for future research based on the approach introduced here.

References

1. Ballard, I., Miller, E.M., Piantadosi, S.T., Goodman, N.D., McClure, S.M.: Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning. Cerebral Cortex **28**(11), 3965–3975 (Nov 2018). https://doi.org/10.1093/cercor/bhx259
2. Chomsky, N.: Three models for the description of language. IRE Transactions on Information Theory **2**(3), 113–124 (Sep 1956). https://doi.org/10.1109/TIT.1956.1056813
3. Enderton, H.B.: A Mathematical Introduction to Logic. Harcourt/Academic Press, San Diego, 2nd ed edn. (2001)
4. Friston, K.J., Lin, M., Frith, C.D., Pezzulo, G., Hobson, J.A., Ondobaka, S.: Active Inference, Curiosity and Insight. Neural Computation **29**(10), 2633–2683 (Oct 2017). https://doi.org/10.1162/neco$_{a0}$0999
5. Goodman, N., Tenenbaum, J., Feldman, J., Griffiths, T.: A Rational Analysis of Rule-Based Concept Learning. Cognitive Science: A Multidisciplinary Journal **32**(1), 108–154 (Jan 2008). https://doi.org/10.1080/03640210701802071
6. Kemp, C., Tenenbaum, J.B., Niyogi, S., Griffiths, T.L.: A probabilistic model of theory formation. Cognition **114**(2), 165–196 (Feb 2010). https://doi.org/10.1016/j.cognition.2009.09.003
7. Levine, S.: Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. arXiv:1805.00909 [cs, stat] (May 2018)
8. Millidge, B., Tschantz, A., Buckley, C.L.: Whence the Expected Free Energy? Neural Computation **33**(2), 447–482 (Feb 2021). https://doi.org/10.1162/neco$_{a0}$1354
9. Piantadosi, S.T., Tenenbaum, J.B., Goodman, N.D.: Bootstrapping in a language of thought: A formal model of numerical concept learning. Cognition **123**(2), 199–217 (May 2012). https://doi.org/10.1016/j.cognition.2011.11.005
10. Piantadosi, S.T., Tenenbaum, J.B., Goodman, N.D.: The logical primitives of thought: Empirical foundations for compositional cognitive models. Psychological Review **123**(4), 392–424 (Jul 2016). https://doi.org/10.1037/a0039980
11. Sipser, M.: Introduction to the Theory of Computation. Boston : PWS Pub. Co. (1997)
12. Smith, R., Schwartenbeck, P., Parr, T., Friston, K.J.: An Active Inference Approach to Modeling Structure Learning: Concept Learning as an Example Case. Frontiers in Computational Neuroscience **14** (2020). https://doi.org/10.3389/fncom.2020.00041
13. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D.: How to Grow a Mind: Statistics, Structure, and Abstraction. Science **331**(6022), 1279–1285 (Mar 2011). https://doi.org/10.1126/science.1192788
14. Ullman, T.D., Goodman, N.D., Tenenbaum, J.B.: Theory Acquisition as Stochastic Search p. 6

Appendix

Context-free grammars

A context-free grammar is defined by a 4-tuple (V, Σ, R, S), with V a finite set of *variables*, Σ a finite set of *terminals*, R a set of *rules*, each of which consist of a variable and a string of variables and terminals, and $S \in V$ is the *start variable*. One can use a grammar to describe a language by generating strings of that language in the following manner.

1. Write down the start symbol.
2. Find a variable that is written down and a rule that starts with that variable. Replace the variable with the right-hand side of the rule.
3. Repeat step 2 until no variables remain.

$$V := \{S, C, D, P, Loc, Col, \wedge, \vee\},$$

Σ is $\{c_1, c_2, c_3, red, green, blue\}$ and rules are as given in section 3. This grammar describes a basic programming language for expressions containing logical connectives and predicates. For example, the string $color(c_1) = red \wedge color(c_2) = green$ can be generated from the grammar. The sequences of substitutions to obtain a string is called a *derivation*. The derivation of the above example is shown in 3.

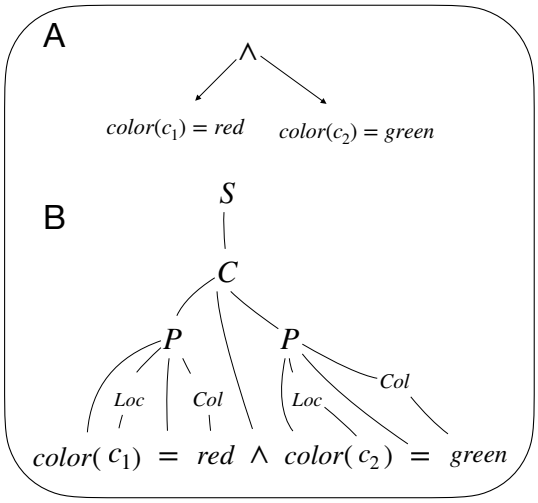


Fig. 3: A: Example of an expression that evaluates to true in the example from 1. B: Corresponding derivation from the language defined by the grammar in section 3.

Interpreting Dynamical Systems as Bayesian Reasoners

Nathaniel Virgo[1][0000−0001−8598−590X], Martin Biehl[2][0000−0002−1670−6855], and Simon McGregor[3]

[1] Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo 152-8550, Japan
nathanielvirgo@elsi.jp
[2] Araya Inc., Tokyo 107-6024, Japan
martin.biehl@gmail.com
[3] University of Sussex, Falmer, UK
s.mcgregor@sussex.ac.uk

Abstract. A central concept in active inference is that the internal states of a physical system parametrise probability measures over states of the external world. These can be seen as an agent's beliefs, expressed as a Bayesian prior or posterior. Here we begin the development of a general theory that would tell us when it is appropriate to interpret states as representing beliefs in this way. We focus on the case in which a system can be interpreted as performing either Bayesian filtering or Bayesian inference. We provide formal definitions of what it means for such an interpretation to exist, using techniques from category theory.

Keywords: Bayesian filtering · Bayesian Inference · Category Theory.

1 Introduction

A question of current interest is *what does it mean for a physical system to be an agent?* That is, given a physical system that interacts with an environment, when does it make sense to say that the system is *learning about its environment* or *trying to achieve a goal*, rather than merely being dynamically coupled to its environment? Here we confine ourselves to the first of these, in a simple form: given a physical system that is influenced by its surroundings, under what circumstances can it be said to be performing inference, such that its internal states could be said to contain 'knowledge' or 'beliefs' about the outside world?

Our approach has something in common with Dennet's intentional stance [17], in that on the one hand we treat the question of whether a system is performing inference as a matter of interpretation, but on the other hand we draw a strong connection between interpretations and the underlying physical dynamics. We provide a formal notion of interpretation for the particular cases we are interested in (Bayesian filtering and Bayesian inference), such that the question of *whether a system can be consistently interpreted in a particular way* is mathematically well-defined and has a definite answer.

The question of how to identify agents in physical systems has been addressed in several ways. Some works focus on whether a system's actions can be seen as pursuing a goal [33,27], with [31] taking an explicitly Dennettian approach. Others focus more on the question of identifying which part of a system should be identified as the agent [7,5,6,28,2], or on understanding what the external world looks like from the agent's point of view [3,5,6]. Another approach, which we take here, is to regard the system's internal state as *parametrising* its beliefs. That is, there is a function mapping the system's physical state to a probability measure that can be seen as a Bayesian prior. This is a key component of work on the Free Energy Principle (FEP) [19,16,34] and also of [39], although our approach differs from these in that our model is not derived from the dynamics of the true environment. The notion that agency is closely related to parametrisation is also central to recent approaches to agency based on category theory [11,38,12].

The idea that states of a system parametrise Bayesian probability distributions appears more broadly in the Bayesian brain literature [26,30] and has also arisen in cell biology [29,32]. Our contribution is to make the concept much more formal, and in so doing, to shed light on the precise relationship between the interpretation level and the underlying physical level.

On a technical level we formulate the problem of Bayesian filtering at an abstract level using the tools of category theory. This part of the work is inspired by [23], which formulates the notion of conjugate prior in terms of category theory in a similar way. Conjugate priors are convenient because they ensure the functional form of the posterior is the same as the posterior, in Bayesian belief updating. At the same time they can be seen as a special case of Bayesian filtering, as we explain in section 2.3 and appendix B.1. Formulating filtering in this way allows us to clearly distinguish the role of the physical machine from the more semantic level at which we can talk about priors and posteriors. We then flip this perspective around, asking, for a given system, whether it can be interpreted as implementing Bayesian filtering, and if so under which model. In this respect our approach generalises that of [8], who studied the special case of the Dirichlet distribution (which is conjugate prior to a categorical distribution) in the context of interpreting a physical system as performing inference.

One thing our approach makes clear is that a given system may have more than one interpretation, and the "correct" interpretation cannot be determined from the system's dynamics alone. Another important aspect of our framework is that an interpretation only depends on the system's internal dynamics, and not on the dynamics of the external world. Because of this, a system's presumed beliefs might not match the true dynamics of the world at all — its beliefs might be consistent but incorrect — and indeed we can construct examples where the world "as the system sees it" has a different causal structure from the world as it really is. (Compare eq. (1) to eq. (7).)

2 Definitions and Results

2.1 Technical preliminaries

In the following, we use the concepts of *measurable space* and *Markov kernel*. By measurable space we mean a set equipped with a σ-algebra, i.e. the kind of thing on which a probability measure can be defined. An example of a measurable space is a finite set, and a reader who is only interested in the finite case could mentally substitute "finite set" wherever we say "measurable space" and "probability distribution" wherever we say "probability measure."

Given a measurable space X, we write $P(X)$ for the set of all probability measures over X. Given measurable spaces X and Y, a Markov kernel is a function $\kappa\colon X \to P(Y)$ that maps elements of X to probability measures over Y, with an additional technical requirement that the function κ be measurable. Markov kernels are closely related to conditional probability, but they are different in that a Markov kernel defines a probability measure over Y for every element X, regardless of whether any probability distribution has been defined over X or what form such a distribution has. In the case where Y is a finite set, we write $\kappa(y \,|\, x)$ for the probability that the kernel κ assigns to y when given the input x.

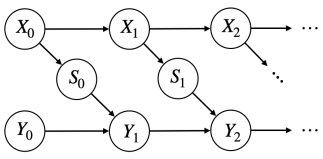
We also make use of a graphical notation known as *string diagrams*, which comes from the literature on category-theoretic probability [13,22]. This notation provides a convenient way to reason about how Markov kernels relate to one another. The full technical description of this calculus can be found in [22] or [13], but we provide a brief intuitive explanation in appendix A, aimed at readers with no category theory background, along with references to further reading.

2.2 Machines and interpretations

We are concerned with interpreting a physical system as performing inferences of some kind on its inputs. We therefore begin by defining a notion corresponding to a physical system that can take an input from the outside world, which leads to a change in state, which might be stochastic and might depend on the input.

Definition 1. *A* machine *consists of two measurable spaces, Y (the state space) and S (the input space), together with a Markov kernel $\gamma\colon Y \times S \to P(Y)$ called the update kernel.*

The idea is that the machine is in reality only a part of some larger stochastic process. We might typically think of this broader context as represented by the following causal Bayesian network, although this is not the only possibility.



(1)

Here the variables $X_0, X_1, \ldots$ represent the states of the external world at different times, which are hidden from the machine's perspective. $S_0, S_1, \ldots$ are the observable "sensor values" that the machine can access, all of which have the same sample space, given by S. Similarly, $Y_0, Y_1, \ldots$ are the machine's internal state at each time step and each has the sample space Y. We assume that each of the nodes $Y_1, Y_2, \ldots$ in this network are associated with the same kernel, γ.

The nodes X_0 and Y_0 represent distributions over initial states of the agent and the world. A common ancestor of these nodes could be added, to represent the possibility that the initial states are correlated. There is no need for any stationarity assumption in our framework; the initial distribution in eq. (1) can be arbitrary.

Even though we might think of the machine as existing in a context along these lines, our notion of interpretation does not depend on the machine's external environment at all, but *only* on the machine's internal dynamics. That is, on the measurable spaces Y and S and on the update kernel γ. This is because, informally speaking, a reasoner may have consistent *but wrong* beliefs about the external world, and we wish to include this possibility in our framework. (In particular, a system that reasons correctly in one environment might be placed in a different environment where the same inferences are no longer correct.) Our notion of interpretation will include a notion of "beliefs" about the external world. These must be consistent with the machine's internal dynamics and the inputs it receives, but they need not relate in any particular way to any ground truth about the process by which those inputs are generated, since we regard that as something the reasoner has no direct access to.

Because of this we will rarely reason about causal models directly, and instead will express our definitions directly in terms of Markov kernels and the relationships between them. The string diagram notation, explained appendix A, will be indispensable for this.

We now describe our central concept: an *interpretation* of a machine. Here we will present only two kinds of interpretations, *Bayesian filtering interpretations* and an important special case, *Bayesian inference interpretations*, but we expect these to fit naturally into a much broader family of concepts.

The most important component of an interpretation is what we term an *interpretation map*, a function that maps the physical state of a machine to something that we can think of as a belief about some external world. In the cases we are concerned with in this paper, a "belief" will be a probability measure over some hidden variable. An interpretation of a machine will be an interpretation map together with some additional data (the *model* as defined below), such that a *consistency equation* is obeyed.

For a given machine there may be many possible interpretations. We use the term *reasoner* for a machine together with a particular choice of interpretation.

In the case of Bayesian inference interpretations and Bayesian filtering interpretations, in which 'beliefs' are probability measures over a hidden variable, the interpretation map is a Markov kernel $\psi_H \colon Y \to P(H)$. Instead of interpreting this stochastically, we think of it as a function that takes a state $y \in Y$ and re-

turns a probability measure $\psi_H(y)$ (the belief) over H. This is to be thought of as the reasoner's subjective knowledge (a Bayesian prior or posterior) about the hypotheses in H. This kernel plays quite a different role from those associated with the graphical model in eq. (1), since its purpose is to map states to beliefs, rather than to model causal influences between random variables.

For an interpretation to be consistent, the reasoner's beliefs must update in the appropriate way when the machine receives new data. The precise meaning of this will depend on what kind of interpretation we are using. For the interpretations we describe here it is given by Bayes' rule, in a form that we state precisely below. In future work we can imagine interpretations based on other principles, such as approximate Bayes (e.g. via the free energy principle).

The idea is that a machine by itself is merely a (possibly stochastic) dynamical process, but if a consistent interpretation exists then it is at least consistent to ascribe a *meaning* to its states.

It should be noted that, for a given machine, the question of *whether it can be consistently interpreted in a particular way* is in principle an empirical one, since it depends on the machine's update kernel, which can in principle be measured. However, in general a given machine might have multiple non-equivalent consistent interpretations, and one cannot distinguish between these empirically by looking only at the system's internal dynamics.[4] Consequently the relationship between interpretations and the empirical, physical world is rather subtle, and one should keep in mind that our notion of "consistent reasoner" unavoidably involves an element of choice in which interpretation to adopt.

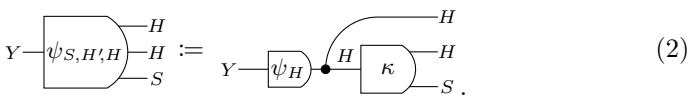
2.3 Consistency for Bayesian Interpretations

We begin with the more general case of Bayesian filtering interpretations. The idea is that a reasoner has a *model* of the environment's dynamics. This model is part of the interpretation, and need not match the true environment dynamics.

In the case of filtering, such a model can be described as a Markov kernel $H \to H \otimes S$, that is, $\kappa\colon H \to P(H \times S)$. The idea is that H is the space of possible hidden states of the external world, as modelled by the reasoner. The kernel κ models a step of this hypothetical external world's evolution, during which it both changes to a new state in H and also emits an observable sensor value in S.

A Bayesian filtering interpretation of a machine $S \otimes Y \to Y$ will then consist of a choice of interpretation map $Y \to H$ as described above, together with a choice of model $H \to H \otimes S$. The kernel κ thus describes a reasoner's beliefs about the next hidden state and the next sensor value, given the current hidden state.

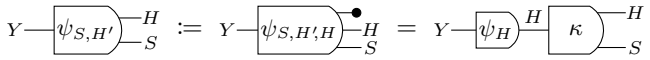
Given the kernels ψ_H and κ we can define another kernel, which we also consider to be an interpretation map,



(2)

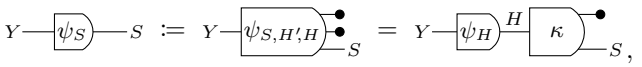
[4] We leave open the possibility that they could be distinguished by looking at some broader context, e.g. by discovering that a device's designer intended a particular interpretation, or that evolution selected for a particular interpretation.

The kernel $\psi_{H,H',S}$ maps a state of the machine $y \in Y$ to a joint distribution over S and two copies of H, which we think of as the reasoner's beliefs about the the next sensor value, the next hidden state, and the current hidden state. We also define its marginals,



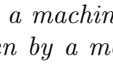
(3)

and



(4)

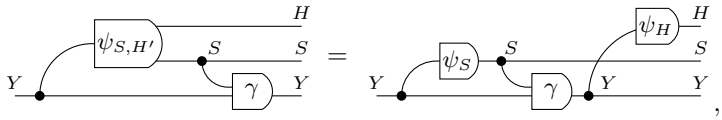
which represent the reasoner's beliefs about the next hidden state and the next input, and about only the next input, respectively. We can now state the consistency requirement for a Bayesian filtering interpretation.

Definition 2. *Given a machine* 

, a consistent Bayesian filtering interpretation *of γ is given by a measurable space H together with Markov kernels*

and

that satisfy



(5)

with $\psi_{S,H'}$ and ψ_S given by eqs. (3) *and* (4)*.*

The left-hand-side of eq. (5) can be read as sampling from the reasoner's joint beliefs about the next hidden state and the next input, and then feeding the corresponding value of S into the machine as an input. The right-hand-side can be read as sampling from the reasoner's belief about its next input, feeding the result in as its next input, and then sampling from its resulting (posterior) belief about what it now sees as the current hidden state. The equation says that these two procedures must give the same result.

In appendix B.1 we give some further intuition for this definition, by considering the case where S, Y and H are finite sets. An important consequence is that, in the finite case, whether a given interpretation is consistent or not only depends on which states are reachable from which other states under a given input; the actual transition probabilities are irrelevant beyond that. We expect an analogous statement to hold more generally.

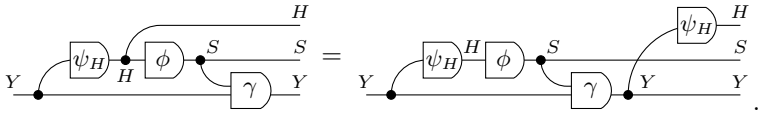
Another important consequence discussed in appendix B.1 is that there is a large class of machines that only admit trivial interpretations. At least in the finite case, non-trivial interpretations can only exist if some transitions are impossible, in the sense that there is a zero probability of transitioning from $y \in Y$ to $y' \in Y$ under the input $s \in S$. So if a finite machine has the property that every state always has a non-zero probability of being the next state, for any given input and initial state, then that machine can only have trivial interpretations.

In appendix B.2 we give a more technical proof, using string diagrams, that at least in the case of deterministic machines, a machine with a consistent Bayesian

filtering interpretation can indeed be regarded as performing a Bayesian filtering task. This can be seen as extending some of the ideas in [24] on conjugate priors to the more general case of Bayesian filtering.

An important special case of definition 2 is where the model κ is such that the hidden state does not change over time. In this case H can be thought of as an unknown parameter of a statistical model, with the sensor inputs being independent and identically distributed samples from the model. We call these *Bayesian inference interpretations*. Although the following follows from definition 2 under this assumption, we write it as a separate definition.

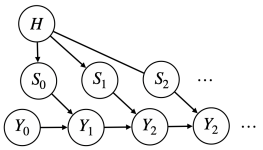
Definition 3. *Given a machine* $\overset{S}{Y}\!\!\rightarrow\boxed{\gamma}\!\!-Y$, *a* consistent Bayesian inference interpretation *of* γ *is given by a measurable space* H *together with Markov kernels* $Y\!-\!\boxed{\psi_H}\!-H$ *and* $H\!-\!\boxed{\phi}\!-S$ *that satisfy*



(6)

In appendix B.3 we show that this definition is closely related to the notion of a conjugate prior, and in particular to the definition of conjugate prior given by [24] in terms of Markov kernels and string diagrams. Finally, in appendix B.4 we unpack eq. (6) in more familiar terms, showing that in the discrete case it does indeed correspond to Bayes' theorem as usually understood.

In a Bayesian inference interpretation we interpret the reasoner as *assuming* its inputs are i.i.d. samples from some distribution, but this need not mean that they actually are. Under a consistent Bayesian inference interpretation a reasoner is interpreted as modelling the world as if its causal structure is as follows, where each of the 'S' nodes is associated with the kernel ϕ.



(7)

However, the true dynamics of the world could still correspond to the Bayesian network in eq. (1) or some other causal structure. The reasoner is simply (interpreted as being) unable to perceive the correlations among its inputs.

In appendix C we present three examples of machines that have consistent Bayesian inference interpretations. The first example is a non-deterministic finite machine with three states. The consistent Bayesian inference interpretation we provide also involves subjectively impossible observations.

The second example is a countably infinite deterministic machine that counts occurrences of each of two possible observations. The consistent Bayesian interpretation we provide intuitively considers this machine to be inferring the bias of a coin that is flipped to cause its observations. It uses the standard conjugate prior to its model as an interpretation map.

The third example is also a countably infinite machine with two possible observations. However, it only stores the difference between the number of the two possible observations. Intuitively, instead of considering all possible biases of a coin as the second example the consistent Bayesian interpretation we provide for this machine infers which of two specific biases of a coin is causing its observations. We also hint at how the second machine can "inherit" this interpretation.

3 Discussion

We see the main contribution of this work as a conceptual one. The consistency equation involved in definitions 2 and 3 can be seen either as a constraint on the machines that have a particular interpretation or as a constraint on the interpretations that a particular machine allows. Every machine has an interpretation with respect to a trivial model (one that has no parameter), but in order to have an interpretation with respect to a non-trivial model, a machine must at least obey the constraints discussed in appendix B.1.

Similar consistency equations should exist for approximate Bayesian filtering as well as for related inference problems like Bayesian smoothing or prediction. Even non-Bayesian normative theories about what a system should be representing and how this should change under external influences will probably have associated consistency equations.

Another direction to extend the concept of consistency equations is to take into account a possible influence of the machine's state on the external world. On the interpretation side this would mean going beyond perception and representation to also include deliberate actions that combine with beliefs and possibly also with goals. A machine with such an interpretation might deserve the term *agent* instead of reasoner.

Our work is related to other current efforts to capture the notion of agent using category theory. These include approaches to Bayesian inference [37] and game theory [10]. The idea that agency is related to parametrisation has also arisen in these contexts [11,38,12]. These works focus on the notion of a *lens* and its generalisations. It is interesting to note that our notions of interpretation seem to be different, and somewhat simpler, in that they lack the bidirectional nature of lenses. We conjecture that a more lens-like bidirectional structure would be needed if we were to consider Bayesian smoothing rather than Bayesian filtering. It will be interesting in future work to better understand the relationship between lens-like categories and the concept of interpretation developed in the present work.

3.1 Relation to the Free Energy Principle

Let us now consider the relation to the Free Energy Principle (FEP) which is also referred to as active inference. The relevant part of FEP is the part called "Bayesian mechanics" [19,16,34]. It seems that the ingredients for an interpretation map can be found in this literature: [16, eq. 3.3] describes a Markov kernel

of an appropriate type, as we detail in appendix D. However, it is not currently clear to us whether the FEP can be formulated in terms of a consistency equation that this kernel obeys. Presumably, such an equation would be different from our definitions 2 and 3, because the FEP is concerned with approximate rather than exact inference and deals with continuous time.

One important difference between our approach and current formulations of FEP is that the FEP requires a stationarity assumption on the true dynamics of the agent-environment system. It seems to us that this is used to derive something that corresponds to a model. In our approach the reasoner's and the "ground truth" dynamics of the environment are different things, and partly for this reason we need no stationarity assumption. We see this conceptual separation as an advantage of the consistency equation approach, and we believe that by incorporating these ideas it might be possible to formulate the FEP in a way that would make its assumptions clearer and perhaps even avoid the need for the stationarity assumption. Although we do not currently know the precise relationship between our work and the FEP at a technical level, we explore it in more detail in appendix D.

Acknowledgements The work by Martin Biehl on this publication was made possible through the support of a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc. Martin Biehl is also funded by the Japan Science and Technology Agency (JST) CREST project.

References

1. Aguilera, M., Millidge, B., Tschantz, A., Buckley, C.L.: How particular is the physics of the Free Energy Principle? arXiv:2105.11203 [q-bio] (May 2021), http://arxiv.org/abs/2105.11203, arXiv: 2105.11203
2. Albantakis, L., Massari, F., Beheler-Amass, M., Tononi, G.: A macro agent and its actions. arXiv:2004.00058 [cs, q-bio] (Mar 2020), http://arxiv.org/abs/2004.00058, arXiv: 2004.00058
3. Ay, N., Löhr, W.: The Umwelt of an embodied agent–a measure-theoretic definition. Theory in Biosciences = Theorie in Den Biowissenschaften **134**(3-4), 105–116 (Dec 2015). https://doi.org/10.1007/s12064-015-0217-3
4. Baez, J., Stay, M.: Physics, Topology, Logic and Computation: A Rosetta Stone. In: Coecke, B. (ed.) New Structures for Physics, pp. 95–172. Lecture Notes in Physics, Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-12821-9_2, https://doi.org/10.1007/978-3-642-12821-9_2
5. Beer, R.D.: Autopoiesis and Cognition in the Game of Life. Artificial Life **10**(3), 309–326 (2004). https://doi.org/10.1162/1064546041255539, /journal/10.1162/1064546041255539
6. Beer, R.D.: The cognitive domain of a glider in the game of life. Artificial Life **20**(2), 183–206 (2014). https://doi.org/10.1162/ARTL_a_00125

7. Biehl, M., Ikegami, T., Polani, D.: Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. In: Proceedings of the Artificial Life Conference 2016. pp. 722–729. The MIT Press (Jul 2016). https://doi.org/10.7551/978-0-262-33936-0-ch115, https://mitpress.mit.edu/sites/default/files/titles/content/conf/alife16/ch115.html
8. Biehl, M., Kanai, R.: Dynamics of a Bayesian Hyperparameter in a Markov Chain. In: Verbelen, T., Lanillos, P., Buckley, C.L., De Boom, C. (eds.) Active Inference. pp. 35–41. Communications in Computer and Information Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-64919-7_5
9. Biehl, M., Pollock, F.A., Kanai, R.: A Technical Critique of Some Parts of the Free Energy Principle. Entropy **23**(3), 293 (Mar 2021). https://doi.org/10.3390/e23030293, https://www.mdpi.com/1099-4300/23/3/293, number: 3 Publisher: Multidisciplinary Digital Publishing Institute
10. Bolt, J., Hedges, J., Zahn, P.: Bayesian open games. arXiv:1910.03656 [cs, math] (Oct 2019), http://arxiv.org/abs/1910.03656, arXiv: 1910.03656
11. Capucci, M., Gavranović, B., Hedges, J., Rischel, E.F.: Towards foundations of categorical cybernetics. arXiv:2105.06332 [math] (May 2021), http://arxiv.org/abs/2105.06332, arXiv: 2105.06332
12. Capucci, M., Ghani, N., Ledent, J., Forsberg, F.N.: Translating Extensive Form Games to Open Games with Agency. arXiv:2105.06763 [cs, math] (May 2021), http://arxiv.org/abs/2105.06763, arXiv: 2105.06763
13. Cho, K., Jacobs, B.: Disintegration and Bayesian Inversion via String Diagrams. Mathematical Structures in Computer Science **29**(7), 938–971 (Aug 2019). https://doi.org/10.1017/S0960129518000488, http://arxiv.org/abs/1709.00322, arXiv: 1709.00322
14. Coecke, B., Paquette, É.: Categories for the Practising Physicist. In: Coecke, B. (ed.) New Structures for Physics, pp. 173–286. Lecture Notes in Physics, Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-12821-9_3, https://doi.org/10.1007/978-3-642-12821-9_3
15. Coecke, B., Kissinger, A.: Picturing quantum processes: a first course in quantum theory and diagrammatic reasoning. Cambridge University Press, Cambridge, United Kingdom ; New York, NY, USA (2017)
16. Da Costa, L., Friston, K., Heins, C., Pavliotis, G.A.: Bayesian Mechanics for Stationary Processes. arXiv:2106.13830 [math-ph, physics:nlin, q-bio] (Jun 2021), http://arxiv.org/abs/2106.13830, arXiv: 2106.13830
17. Dennett, D.C.: True Believers : The Intentional Strategy and Why It Works. In: Heath, A.F. (ed.) Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford, pp. 53–75. Clarendon Press (1981)
18. Fong, B., Spivak, D.I.: An invitation to applied category theory: seven sketches in compositionality. Cambridge University Press, Cambridge ; New York, NY (2019)
19. Friston, K.: A free energy principle for a particular physics. arXiv:1906.10184 [q-bio] (Jun 2019), http://arxiv.org/abs/1906.10184, arXiv: 1906.10184
20. Friston, K., Da Costa, L., Hafner, D., Hesp, C., Parr, T.: Sophisticated Inference. Neural Computation **33**(3), 713–763 (Mar 2021). https://doi.org/10.1162/neco_a_01351, https://doi.org/10.1162/neco_a_01351
21. Friston, K.J., Da Costa, L., Parr, T.: Some Interesting Observations on the Free Energy Principle. Entropy **23**(8), 1076 (Aug 2021). https://doi.org/10.3390/e23081076, https://www.mdpi.com/1099-4300/23/8/1076, number: 8 Publisher: Multidisciplinary Digital Publishing Institute

22. Fritz, T.: A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. Advances in Mathematics **370**, 107239 (Aug 2020). https://doi.org/10.1016/j.aim.2020.107239, https://www.sciencedirect.com/science/article/pii/S0001870820302656
23. Jacobs, B.: A channel-based perspective on conjugate priors. Mathematical Structures in Computer Science **30**(1), 44–61 (Jan 2020). https://doi.org/10.1017/S0960129519000082, https://www.cambridge.org/core/journals/mathematical-structures-in-computer-science/article/channelbased-perspective-on-conjugate-priors/D7897ABA1AA06E5F586F60CB21BDDB32, publisher: Cambridge University Press
24. Jacobs, B.: A Channel-Based Perspective on Conjugate Priors. arXiv:1707.00269 [cs] (Sep 2018), http://arxiv.org/abs/1707.00269, arXiv: 1707.00269
25. Jacobs, B., Staton, S.: De Finetti's Construction as a Categorical Limit. In: Petrişan, D., Rot, J. (eds.) Coalgebraic Methods in Computer Science. pp. 90–111. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-57201-3_6
26. Knill, D.C., Pouget, A.: The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences **27**(12), 712–719 (Dec 2004). https://doi.org/10.1016/j.tins.2004.10.007, https://www.cell.com/trends/neurosciences/abstract/S0166-2236(04)00335-2, publisher: Elsevier
27. Kolchinsky, A., Wolpert, D.H.: Semantic information, autonomous agency and non-equilibrium statistical physics. Interface Focus **8**(6), 20180041 (Dec 2018). https://doi.org/10.1098/rsfs.2018.0041, https://royalsocietypublishing.org/doi/full/10.1098/rsfs.2018.0041
28. Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J.C., Ay, N.: The information theory of individuality. Theory in Biosciences **139**(2), 209–223 (Jun 2020). https://doi.org/10.1007/s12064-020-00313-7, https://doi.org/10.1007/s12064-020-00313-7
29. Libby, E., Perkins, T.J., Swain, P.S.: Noisy information processing through transcriptional regulation. Proceedings of the National Academy of Sciences **104**(17), 7151–7156 (Apr 2007)
30. Ma, W.J., Jazayeri, M.: Neural coding of uncertainty and probability. Annual Review of Neuroscience **37**, 205–220 (2014). https://doi.org/10.1146/annurev-neuro-071013-014017
31. McGregor, S.: The Bayesian stance: Equations for 'as-if' sensorimotor agency. Adaptive Behavior p. 105971231770050 (Mar 2017). https://doi.org/10.1177/1059712317700501, http://journals.sagepub.com/doi/10.1177/1059712317700501
32. Nakamura, K., Kobayashi, T.J.: Connection between the Bacterial Chemotactic Network and Optimal Filtering. Physical Review Letters **126**(12), 128102 (Mar 2021). https://doi.org/10.1103/PhysRevLett.126.128102, https://link.aps.org/doi/10.1103/PhysRevLett.126.128102
33. Orseau, L., McGill, S.M., Legg, S.: Agents and Devices: A Relative Definition of Agency. arXiv:1805.12387 [cs, stat] (May 2018), http://arxiv.org/abs/1805.12387, arXiv: 1805.12387
34. Parr, T., Da Costa, L., Friston, K.: Markov blankets, information geometry and stochastic thermodynamics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **378**(2164), 20190159 (Feb 2020). https://doi.org/10.1098/rsta.2019.0159, https://royalsocietypublishing.org/doi/full/10.1098/rsta.2019.0159

35. Risken, H., Frank, T.: The Fokker-Planck Equation: Methods of Solution and Applications. Springer Series in Synergetics, Springer-Verlag, Berlin Heidelberg, 2 edn. (1996). https://doi.org/10.1007/978-3-642-61544-3, https://www.springer.com/gp/book/9783540615309
36. Rosas, F.E., Mediano, P.A.M., Biehl, M., Chandaria, S., Polani, D.: Causal Blankets: Theory and Algorithmic Framework. In: Verbelen, T., Lanillos, P., Buckley, C.L., De Boom, C. (eds.) Active Inference. pp. 187–198. Communications in Computer and Information Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-64919-7_19
37. Smithe, T.S.C.: Bayesian Updates Compose Optically. arXiv:2006.01631 [math, stat] (Jul 2020), http://arxiv.org/abs/2006.01631, arXiv: 2006.01631
38. St Clere Smithe, T.: Cyber Kittens, or Some First Steps Towards Categorical Cybernetics. Electronic Proceedings in Theoretical Computer Science **333**, 108–124 (Feb 2021). https://doi.org/10.4204/EPTCS.333.8, http://arxiv.org/abs/2101.10483v1
39. Still, S., Sivak, D.A., Bell, A.J., Crooks, G.E.: The thermodynamics of prediction. arXiv e-print 1203.3271 (Mar 2012), http://arxiv.org/abs/1203.3271, phys. Rev. Lett. 109, 120604 (2012)
40. Wikipedia contributors: Conjugate prior — Wikipedia, the free encyclopedia (2021), https://en.wikipedia.org/w/index.php?title=Conjugate_prior&oldid=1030202570, [Online; accessed 8-July-2021]

A Category-Theoretic Probability and String Diagrams

In this paper we use some concepts from category-theoretic probability, and in particular we use a notation known as string diagrams. A full introduction to these topics would be out of scope of the paper, but we include here an informal introduction to the topic. We do this because, to our knowledge, no concise introduction currently exists that is focused on (classical) probability and does not assume a background in category theory. We assume that the reader knows the definition of a category, but not much more than that.

Appendix A.1 introduces the basic concepts, mostly in the context of discrete probability. In appendix A.2 we briefly comment on how this extends to the general case of measure-theoretic probability with very little extra work. In appendix A.3 we explain how to reason about conditional probabilities and Bayes' theorem within this category-theoretic context.

These sections contain no original material. Their purpose is to give the reader enough information to be able to read the string diagram equations in the main text and later sections of the appendix without needing to consult a category theory text. However, they are intended neither as an authoritative technical reference nor as a comprehensive review, and readers should consult the cited references for full details.

A.1 Introduction to String Diagrams and Category-Theoretic Probability

A full technical introduction to the use of string diagrams in probability can be found in [22] or the earlier [13], but these works require some knowledge of

category theory. The string diagram notation predates its use in probability and has many other applications. One could consult [4,14,18,15] for tutorial introductions to diagrammatic reasoning in other fields, of various different flavours. Here we present it somewhat informally and only in the context of probability.

It should be kept in mind that, despite our somewhat informal introduction, string diagrams are formal expressions. The main difference between them and the more familiar kind of mathematical expression formed from strings of symbols is their two-dimensional syntax. This makes it easier to express certain concepts. (Particularly those relating to joint distributions, in the case of probability.)

We use the so-called *Markov category* approach to probability [22]. The main idea here is to express everything in terms of *measurable spaces* and *Markov kernels*, whose definitions we outlined in the main text.[5] To explain how the framework works, let us consider the special case where the only measurable spaces we are interested in are finite sets (with their power sets as their σ-algebras). If A and B are finite sets then a Markov kernel can be thought of as just a function $f\colon A \to P(B)$, where $P(B)$ is the set of all probability distributions over B. (The set $P(B)$ may be thought of as a $(|B|-1)$-dimensional simplex, consisting of all those vectors in $\mathbb{R}^{|B|}$ whose components are all non-negative and sum to 1.) Such a function amounts to a $|B|$-by-$|A|$ stochastic matrix, although some care needs to be taken over which rows correspond to which elements of B and which columns to which elements of A.

In this finite case, we write $f(b \,|\, a)$ to denote the probability that the kernel f assigns to the outcome $b \in B$ when given the input $a \in A$. We use a thick vertical line to indicate a close relationship to conditional probability while also emphasising that the concept is different: given a kernel $f\colon A \to P(B)$ the quantities $f(b \,|\, a)$ are always defined, regardless of whether any probability distribution has been defined over A, and regardless of whether a has a nonzero probability according to such a distribution. More common notations include $|$ or $;$ in place of $|$.

We also write $f(a)$ for the probability distribution over B that the function f returns when given the input a. We could say that $f(b \,|\, a)$ is defined as $f(a)(b)$.

Given Markov kernels $f\colon A \to P(B)$ and $g\colon B \to P(C)$, we can compose them to form a new kernel of type $A \to P(C)$. We write this $f \fatsemi g$. It is given by

$$(f \fatsemi g)(c \,|\, a) = \sum_{b \in BY} f(b \,|\, a)\, g(c \,|\, b). \tag{8}$$

In this finite case this is simply matrix multiplication, and we could have denoted it gf instead of $f \fatsemi g$ accordingly. (Another common notation is $g \circ f$.) We prefer

[5] In fact for most of the paper we will work much more abstractly than this. It would be more correct to say "objects in a Markov category" wherever we say "measurable space" and "morphisms in a Markov category" wherever we say "Markov kernel," since for most of the paper we will reason at the category level, and we will not directly invoke the definition of a measurable space. We have chosen to use the more concrete terms because they express a clear intuition for how these objects and morphisms are intended to be interpreted.

$f \fatsemi g$ because it puts f and g in the same order that they will appear in string diagrams.

It is straightforward to show that composition is associative, that is

$$(f \fatsemi g) \fatsemi h = f \fatsemi (g \fatsemi h). \tag{9}$$

In addition, for every finite set A there is an identity kernel, which amounts to just the $|A|$-by-$|A|$ identity matrix. We write this as id_A and define it by $\mathsf{id}_A(a' \,|\, a) = \delta_{a,a'}$. For every Markov kernel $f \colon A \to P(B)$ we have

$$\mathsf{id}_A \fatsemi f = f = f \fatsemi \mathsf{id}_B. \tag{10}$$

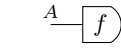
These two facts mean that there is a category whose objects are finite sets and whose morphisms are Markov kernels between finite sets. This category is called $\mathsf{FinStoch}$.

Since Markov kernels are morphisms in a category, we will often write $f \colon A \kto B$ instead of $f \colon A \to P(B)$, using the dotted arrow $\kto$ to distinguish morphisms in $\mathsf{FinStoch}$ and related categories from ordinary functions. (In the main text we continue writing them as functions in order to avoid introducing new notation.)

The composition of Markov kernels can be generalised to the case of measure-theoretic probability, which allows us to reason about continuous probability and more general probability measures using the same kinds of diagram and much of the same reasoning. We briefly discuss this in more detail in appendix A.2. The main difference is that composition becomes integration over measures rather than summation.

Probability measures themselves may be seen as a special case of Markov kernels. Consider a set with a single element, denoted $\mathbb{1} = \{\star\}$. (The identity of the element does not matter because all one-element sets are isomorphic to each other. Category theorists often speak of "the one-element set" for this reason. We use a star to denote the element.) Then a Markov kernel $p \colon \mathbb{1} \kto A$ is a function $p \colon \mathbb{1} \to P(A)$, which takes an element of $\mathbb{1}$ and returns a probability measure over A. Since there is only one element of $\mathbb{1}$ this means that the kernel p only defines a single probability measure over A. We therefore think of Markov kernels $\mathbb{1} \kto A$ and probability measures over A as essentially the same thing.

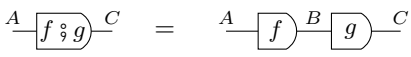
We now begin to introduce the string diagram notation. A Markov kernel $f \colon A \kto B$ will be denoted



. (11)

This expression means much the same thing as the notation $f \colon A \kto B$. It is just a formal symbol denoting the kernel f, annotated with type information.

The composition of kernels $f \colon A \kto B$ and $g \colon B \kto C$ is written



. (12)

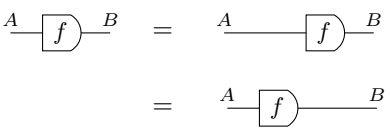
The left and right hand side of this equation are just two different ways to write the composite kernel $f \fatsemi g$, as defined by eq. (8) or its measure-theoretic generalisation.

In reading a diagram like the right-hand side of eq. (12) we find it helpful to imagine an element of A travelling along the wire from the left. As it passes through the kernel f it is stochastically transformed into an element of B, in a way that might depend on its original value. It then travels further to the right and is stochastically transformed by g into an element of C. Equation (8) can be seen as describing this process.

In string diagrams a special notation is used for identity kernels (or identity morphisms more generally): an identity kernel id_A is drawn simply as a wire with no box on it,

. (13)

For any Markov kernel $f\colon A \to B$ the identity law eq. (10) can then be written



. (14)

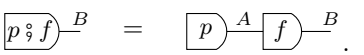
This allows us to think of the wires as stretchy: we can extend and contract them at will. We will think of the wires as continuously deformable, rather than extending and contracting in discrete units. This is justified by the formal theory of string diagrams. (One may informally think of the wire itself as an infinite chain of identity kernels, all composed together.) This ability to continuously deform diagrams turns out to be an extremely powerful and useful idea.

Another special notation is used for one-element sets[6]: they are drawn as no wire at all. For this reason a probability measure over A, that is, a kernel $p\colon \mathbb{1} \to A$, is drawn as

. (15)

(Morphisms of this kind are sometimes known as "states," and they are often drawn as a triangle rather than a box, though here we draw them in the same style as other morphisms.)

It is worth noting that the kernels p and f above can be composed, yielding



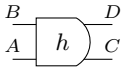
. (16)

Because of this, although the kernel $f\colon A \to B$ is defined as a function $f\colon A \to P(B)$ mapping *elements* of A to probability distributions over B, we can instead choose to see it as mapping *probability measures* over A to probability measures over B. In the finite case, if we think of finite probability distributions as normalised and nonnegative vectors in $\mathbb{R}^n$, then f can be seen as a linear map with the property that it maps points in one simplex to points in another. (This justifies thinking of it as a stochastic matrix.)

The string diagram notation becomes useful when we start thinking about joint distributions. We do this by drawing wires in parallel. As an example, we can consider a Markov kernel defined by a function $h\colon A \times B \to P(C \times D)$.

[6] or in a more general context, the unit object of a monoidal category

This function takes two arguments, an element of A and an element of B, and it returns a joint probability distribution over C and D. In string diagrams we write this as



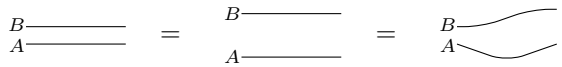
(17)

In symbols, we write $h\colon A \otimes B \to C \otimes D$. An object like $A \otimes B$, drawn as two parallel wires, can either be thought of as the measurable space $A \times B$ (which is the Cartesian product of sets in the finite case), or as the space of probability measures over $A \times B$. The symbol $\otimes$ is referred to as a monoidal product.

There is some inherent ambiguity in this notation. If we draw three parallel wires, 

, it could either mean $(A \otimes B) \otimes C$ or $A \otimes (B \otimes C)$. In the finite case, these correspond to the sets $(A \times B) \times C$ and $A \times (B \times C)$. These are different sets, since one is composed of pairs $((a, b), c)$ and the other of pairs $(a, (b, c))$. This ambiguity is not important in practice, however, and the formal machinery of *monoidal categories* allows us to use string diagrams without worrying about it. We do not give a formal treatment of this here. (A concise summary can be found in [4].) Instead we simply remark that when we draw three parallel wires we think of joint distributions over A, B and C, and the precise distinction between $P(A \times (B \times C))$ and $P((A \times B) \times C)$ will not be important to us.

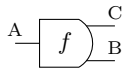
In a similar vein, the spaces A and $A \otimes \mathbb{1}$ are different, but the difference is not important to us, and in fact they are written the same way in string diagrams. This is because we draw $\mathbb{1}$ as an invisible wire. This also allows us to write



(18)

That is, string diagrams are stretchy in the vertical direction as well as the horizontal one. We can bend the wires, as long as we don't deform them so much that they point backwards, from right to left.

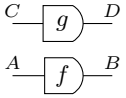
This also allows to write things like



(19)

for a kernel $f\colon A \to B \otimes C$.

We can also draw morphisms (i.e. Markov kernels) in parallel with each other, for example,



(20)

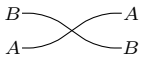
We write this in symbols as $f \otimes g$, which is a morphism of type $A \otimes C \to B \otimes D$. In the finite case, it is given by

$$(f \otimes g)(b, d \,|\, a, c) = f(b \,|\, a)\, g(d \,|\, c). \tag{21}$$

The probabilities $f(b \,|\, c)$ and $g(d \,|\, c)$ are multiplied together because the two Markov kernels are operating in parallel. One can imagine an element of A entering from the bottom left and being stochastically transformed by f into an

element of B, while in parallel, and independently, an element of C enters from the top left and is stochastically transformed by g into an element of D. In general, in the finite case, $f \otimes g$ is given by the tensor product of the stochastic matrices that represent f and g. (This might give some intuition for the symbol $\otimes$.)

We can cross wires over each other. (In category theory terms, the categories we are concerned with are symmetric monoidal categories.) The diagram



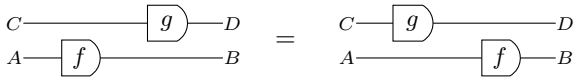
(22)

can be seen as a Markov kernel $A \otimes B \to B \otimes A$. In the finite case it is defined by

$$\mathrm{swap}_{A,B}(b', a' \,|\, a, b) = \delta_{a,a'}\delta_{b,b'}. \tag{23}$$

We have a number of equations that are standard in monoidal category theory, and allow us to freely slide boxes along wires and bend wires to cross over each other. These can either be shown directly from the definitions above or (perhaps more usefully) deduced from the definition of a symmetric monoidal category. Three such equations are as follows. More details can be found in the references cited above.

(24)



(25)

(26)

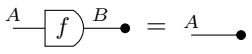
So far, everything we have said about string diagrams applies to any symmetric monoidal category. However, there are two additional things we can add that take us much closer to probability theory. These are the ability to *copy* and to *delete*. These operations, and their special properties, do not necessarily exist in other contexts, such as quantum mechanics. This is a central point of [4,14]. We will stick to the context of classical probability, however, so copying and deletion will always be possible in this paper.

We cover deletion first. For every measurable space A there is a unique kernel of type $A \to \mathbb{1}$, which we call del_A. In the finite case it is given by $\mathsf{del}_A(\star \,|\, a) = 1$ for all $a \in A$. We can think of this as a $1 \times |A|$ matrix (i.e. a row vector) whose entries are all 1. This is the only possible $1 \times |A|$ stochastic matrix.

In string diagrams we write such a deletion kernel as a black dot:

. (27)

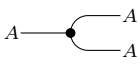
There is one such morphism for every measurable space, but we denote them all with the same kind of black dot. These black dots have the property that



(28)

for every Markov kernel f. This says that if we take some input A, perform some stochastic operation f on it and then delete the result, this is the same as simply deleting the input.[7]

The second special operation is copying. For every measurable space A there is a kernel $\mathsf{copy}_A \colon A \to A \otimes A$, which we will describe shortly. We write this also as a black dot, but this time with two output wires rather than one.



(29)

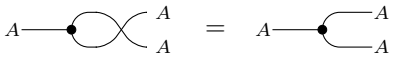
Informally, this kernel takes an outcome $a \in A$ and copies it, producing a pair (a, a) of identical values. It's important to note that it copies *values* rather than *distributions*. Its output does not consist of two independent and identically distributed elements of A but rather two perfectly correlated elements of A that always have the same value. In the discrete case the copy map is defined as

$$\mathsf{copy}_A(a'', a' \,|\, a) = \begin{cases} 1 & \text{if } a'' = a' = a \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

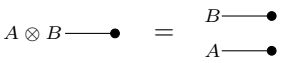
In addition to eq. (28), the copy and delete maps obey the following properties [22, definition 2.1]:

(31)

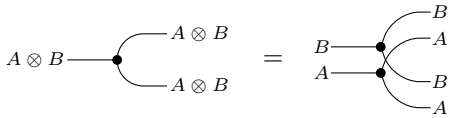
(32)



(33)



(34)



(35)

Equation (31) says that if we make multiple copies of something it doesn't matter which order we make them in. Equation (32) says that if we copy something and then delete one of the copies, that is the same as doing nothing to it. Equation (33) says that if we copy something and then swap the copies it makes no difference. (Because the two copies are the same as each other.)

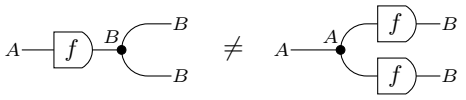
[7] In category theory terms, this means that the set of all delete kernels collectively forms a natural transformation. (Specifically, it is a natural transformation from the identity functor to the functor that sends all objects to $\mathbb{1}$ and all morphisms to $\mathsf{id}_{\mathbb{1}}$.) For this reason this property of delete kernels is called "naturality."

Equations (34) and (35) are more technical. They say that if we have elements of A and B we can delete or copy them as a single element of $A \otimes B$ or separately, as elements of A and B, and these should give the same result.

These equations can be derived from the definitions we have given for the finite case. They may also be derived in various more general measure-theoretic contexts [13,22].

However, the approach of [22] is instead to treat them as *axioms*: any symmetric monoidal category with copy and delete maps that obey eqs. (28) and (31) to (35) is called a *Markov category*. One can do a surprising amount of reasoning about probability theory using these axioms alone, although there are also Markov categories that do not directly resemble the category of measurable spaces and Markov kernels that we have described. There are various additional axioms that can to be added as well, which then allow more specific results to be proven. (See [22] for the details.)

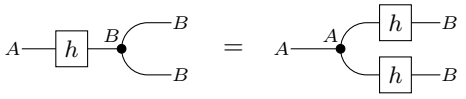
An important thing to note about the copy operator is that, in general,



(36)

That is, copying the output of a kernel f is not the same as copying its input and then applying two copies of the kernel to it. Intuitively, this is because f might be stochastic. If we copy the output we end up with two perfectly correlated copies, whereas if we copy the input then the stochastic variations will be independent.

However, if the kernel is deterministic then copying its input is indeed the same as copying its output. In fact, in the Markov category framework this is the *definition* of a deterministic Markov kernel: we say a kernel $h\colon A \to B$ is deterministic if



(37)

In this paper we use square boxes for kernels that are known to be deterministic, and boxes with rounded edges for general, possibly-stochastic kernels.

In the main text, we write Markov kernels as functions $f\colon A \to P(B)$, and we write deterministic kernels as functions $f\colon A \to B$. To be more precise, a deterministic kernel should really also be considered as a function $f\colon A \to P(B)$, such that eq. (37) is obeyed. However, if we assume we are working in a category called BorelStoch (which is a common assumption in category-theoretic probability) then eq. (37) implies that f always returns a delta measure [22, example 10.5], and in this case there is not much harm in treating a deterministic kernel f as a function $f\colon A \to B$.

A.2 The extension to measure theory

Above we described the category FinStoch and introduced string diagrams mostly in that context. Here we briefly describe how this generalises to the measure-theoretic case, which is needed in order to think about continuous probability.

In the measure-theoretic case the objects (X, Y, etc.) are any measurable spaces rather than only finite sets. Markov kernels are still functions $f\colon X \to P(Y)$, but now $P(Y)$ is the set of all probability measures on the measurable space Y. (That is, $P(Y)$ is the set of all functions from the σ-algebra associated with Y to $[0,1]$, such that Kolmogorov's axioms are obeyed.) $P(Y)$ can itself be made into a measurable space in a standard way, and the function f must obey an additional restriction that it be a measurable function. (This means that the preimage of every element of $P(Y)$ must be a member of the σ-algebra associated with X.)

In this case $f(x)$ is a probability measure rather than a probability distribution, and composition is given by integration rather than summation. (See [22, example 4] for the details.) This gives rise to a category called Stoch, whose objects are all measurable spaces and whose morphisms are all Markov kernels. (This category is also known as the Kleisli category of the Giry monad, for reasons we do not discuss here.)

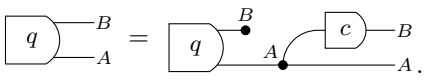
Unfortunately the category Stoch does not have all of the properties that one might want it to have. (See appendix A.3 below.) Because of this a common approach is to work in a category called $\mathsf{BorelStoch}$ (also discussed in [22, example 4]), in which the objects are a subset of measurable spaces called standard Borel spaces, and the morphisms are all Markov kernels between standard Borel spaces. Standard Borel spaces include many kinds of measurable space that one would be likely to use in practice, and in particular they include both finite sets and $\mathbb{R}^n$ with its usual σ-algebra.

In the present paper, the properties of $\mathsf{BorelStoch}$ are used in two ways. Firstly, in $\mathsf{BorelStoch}$ we can always use conditionals, as explained in the next section. Secondly, as a notational convenience we treat deterministic kernels and measurable functions as interchangeable, which makes sense in $\mathsf{BorelStoch}$ but doesn't hold in the more general case of Stoch.

A.3 Conditionals and Bayes' theorem

Conditional probabilities and Bayes' theorem play central roles in the theory of inference. Here we briefly discuss how they look in string diagrams. Given a joint distribution $\boxed{q}\!\!\begin{smallmatrix}B\\A\end{smallmatrix}$ we may want to split it up into a product of a marginal and a conditional, which in traditional notation, in the discrete case, would be written $p(a,b) = p(a)\,p(b\,|\,a)$.

The category-theoretic approach, as set out in [13,22], is slightly different. We write the following, which is called a *disintegration* of q. (The term "disintegration" is used because it is the opposite of integration.)



(38)

Here, $\boxed{q}\!\!-\!\!\bullet_A$ is the marginal of A according to the joint distribution q. In the finite case it can be written $\sum_{b\in B} q(a,b)$. The kernel $A\text{—}\boxed{c}\text{—}B$ is called a *conditional*

of p. It is defined by eq. (38), which in the finite case can be written

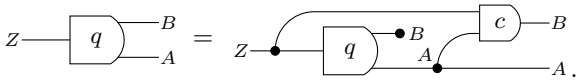
$$q(a,b) = \left(\sum_{b' \in B} q(a,b')\right) c(b \,|\, a). \tag{39}$$

This is closely analogous to the identity $p(a,b) = p(a)\, p(b \,|\, a)$. The difference is that $p(b \,|\, a)$ is defined as $p(a,b)/p(a)$, and is only defined when $p(a) > 0$. On the other hand, in eq. (39), if $\left(\sum_{b' \in B} q(a,b')\right) = 0$ for some $a \in A$ then $q(a,b)$ must be 0 for all $b \in B$, and consequently the equation puts no constraint on $c(b \,|\, a)$ in this case.

This means that instead of being undefined in this case, the conditional c is not *uniquely* defined: there may be many different kernels c that satisfy the equation.

This carries over to the general measure-theoretic case as well. If we are in the category BorelStoch then for any joint distribution $\boxed{q}\!\!\begin{smallmatrix}B\\A\end{smallmatrix}$ there exists at least one conditional $A\!-\!\boxed{c}\!-\!B$ that satisfies eq. (38), but there might be many. (In the case of Stoch conditionals may fail to exist at all, see [22, example 11.3].)

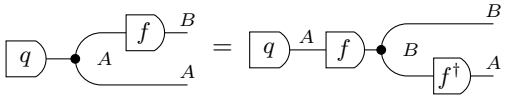
We may also want to disintegrate a joint distribution that is a function of some parameter, e.g. $Z\!-\!\boxed{q}\!\!\begin{smallmatrix}B\\A\end{smallmatrix}$. In this case eq. (38) becomes



(40)

Conceptually this is very similar. We want the disintegration to hold for every parameter value $z \in Z$, and we define the conditional to be a function of z as well as of $a \in A$. In the discrete case, eq. (40) is analogous to the identity $p(a,b \,|\, z) = p(a \,|\, z)\, p(b \,|\, a,z)$.

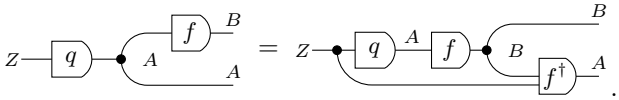
Bayes' theorem is closely related to conditional probability and can be expressed in a similar way. Given a prior $\boxed{q}\!-\!A$ and a kernel $A\!-\!\boxed{f}\!-\!B$, we can define a *Bayesian inverse* of f with respect to q, which is a kernel $B\!-\!\boxed{f^\dagger}\!-\!A$ such that



(41)

The Bayesian inverse $f^\dagger$ depends on the prior q as well as on the kernel f. If we had chosen a different distribution in place of q, the Bayesian inverse $f^\dagger$ would be different. As with conditionals, Bayesian inverses are not necessarily unique, and for a given f and q there may be many kernels $f^\dagger$ that satisfy eq. (41). (In fact, Bayesian inverses can be seen as a special case of conditionals; see [13,22].)

We may also consider the case where the prior takes a parameter, such as $Z\!-\!\boxed{q}\!-\!A$. In this case a Bayesian inverse also needs to depend on the parameter in general, which gives us the following more general definition:



(42)

The references [13,22,37] contain much more detail about Bayes' theorem in this form.

B More details about Bayesian interpretations

B.1 Unpacking Bayesian filtering interpretations

In this section we give some more intuition for definition 2 and then note some consequences of it. The section deals mostly with the case where S, Y and H are discrete sets, meaning that we can reason in terms of probability distributions rather than measure theory. In this case definition 2 can be written in a form that makes the relationship to Bayes' theorem more clear. We define a notion of *subjectively impossible input*, which is a value of S that the reasoner believes with certainty will not occur as its next input. (This does not imply that the input actually is impossible according to the true dynamics of the environment.) We show that definition 2 puts no constraints on the reasoner's posterior after receiving a subjectively impossible input.

We also show that the possible interpretations of a machine only depend on which states can transition to which other states given which inputs, and not on the probabilities of such transitions. In addition, we show that some machines admit no non-trivial interpretations at all.

In order to unpack definition 2 a little more, let us consider the case where S, Y and H are discrete. Before starting we note that in the finite case, the definition of ψ_S, eq. (4), can be written as

$$\psi_S(s \,|\, y) = \sum_{h \in H} \psi_{S,H'}(s, h \,|\, y). \tag{43}$$

In this case, eq. (5) can be written in symbols as

$$\psi_{S,H'}(h, s \,|\, y)\gamma(y' \,|\, y, s) = \psi_S(s \,|\, y)\gamma(y' \,|\, y, s)\psi_H(h \,|\, y'), \tag{44}$$

for all $s \in S, h, \in H, y, y' \in Y$. We can cancel $\gamma(y' \,|\, y, s)$ from both sides on the assumption that it is positive, yielding

$$\gamma(y' \,|\, y, s) > 0 \implies \psi_{S,H'}(h, s \,|\, y) = \psi_S(s \,|\, y)\psi_H(h \,|\, y'). \tag{45}$$

The condition $\gamma(y' \,|\, y, s) > 0$ means that $y' \in Y$ is a *possible next state* when the machine starts in state $y \in Y$ and receives the input $s \in S$. (There may be many possible next states in this situation because the machine may be stochastic.)

Let us then suppose that the machine starts in state y, receives an input s, and transitions to state y'. Let h be an arbitrary element of H. The number $\psi_{S,H'}(h, s \,|\, y) \in [0, 1]$ can then be seen as the reasoner's prior probability that the next state is h and the next input is s. In more traditional notation we might write this as $P(H' = h, S = s)$, where we leave the state of the underlying

machine implicit. (Here we do not attempt to formalise this in terms of random variables, but simply treat it as a kind of notational shorthand for $\psi_{S,H'}(h, s \,|\, y)$.)

We may then regard $\psi_S(s \,|\, y)$ as the reasoner's prior probability that the next input is s, i.e. $P(S = s) = \sum_{h \in H} P(H = h, S = s)$.

However, since $\psi_H(h \,|\, y')$ is conditioned on y' rather than y, we instead regard it as the reasoner's *posterior* probability that $H' = h$. (We refer to H' rather than H here because after it receives an input its previous "next" hidden state becomes its current hidden state.) $\psi_H(h \,|\, y')$ therefore corresponds to what we might write as $P(H' = h \mid S = s)$.

With this informal shorthand notation eq. (45) then says

$$P(H' = h, S = s) = P(S = s)\, P(H' = h \mid S = s), \tag{46}$$

which has the same appearance as a familiar identity from elementary probability theory. It corresponds to a single step of Bayesian filtering, which we spell out in more detail in appendix B.2.

This shorthand notation gives some intuition for why eq. (5) has the particular form it does, but it leaves the dependence on the state of the underlying machine implicit, and in so doing it obscures an important and subtle point. In a more traditional context, $P(H' = h \mid S = s)$ is defined by

$$P(H' = h \mid S = s) = P(H' = h, S = s)/P(S = s) \tag{47}$$

and has no value when $P(S = s) = 0$. However, in our case $P(H' = h \mid S = s)$ is a shorthand for $\psi_H(h \,|\, y')$, which is defined even when $\psi_S(s \,|\, y) = 0$.

In the case where $P(S = s) > 0$, eq. (5) in the form of eq. (46) demands that $P(H' = h \mid S = s)$ is indeed equal to $P(H' = h, S = s)/P(S = s)$. More precisely, if $\psi_S(s \,|\, y) > 0$ then we must have $\psi_H(h \,|\, y') = \psi_{S,H'}(h, s \,|\, y)/\psi_S(s \,|\, y)$. However, if $\psi_S(s \,|\, y) = 0$ then eq. (5) puts no constraints on $\psi_{S,H'}(h, s \,|\, y)$ at all, or indeed on $\psi_H(h \,|\, y)$.

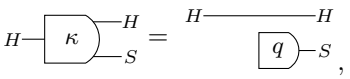
In the case where S is a discrete set (even if Y and H are not discrete), we say that $s \in S$ is a *subjectively impossible input* for a given state $y \in Y$ if $\psi_S(s \,|\, y) = 0$. The point is that the reasoner believes, with certainty, that it will not receive the input s as its next input. The reasoning above says that in this situation, *any* posterior over H is acceptable, because Bayes' rule doesn't specify what the posterior should be. We find this somewhat analogous to the fact that in logic one can deduce any proposition from a contradiction. Definition 2 indeed permits any posterior in the case of a subjectively impossible input. In fact, it even allows the posterior to be chosen stochastically in this case.

This is in a sense the minimal possible assumption we could make. However, one could imagine addressing the issue in a different way by changing the framework, thus introducing a subtly different notion of interpretation than the one we have presented here. One possibility would be to allow *partial interpretations*, where ψ_H becomes a partial function, meaning that not every state of the machine needs to have an interpretation at all. This would allow the posterior to be undefined in the case of a subjectively impossible input, rather than merely arbitrarily defined. Another possibility would be to strengthen eq. (5)

with additional conditions, forcing the posterior to be meaningful even after a subjectively impossible input. We suspect that such an approach can lead to an interesting way to formalise improper priors, which are also about having meaningful posteriors in the case of 'impossible' inputs, but we leave investigation of this to future work.

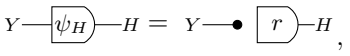
We note one other important consequence of the above reasoning, in the discrete case. When we express eq. (5) in the form of eq. (45), we see that it only depends on whether a transition from y to y' is possible given an input s, and not on the probability of such a transition. Thus, for Bayesian filtering interpretations (and hence also for Bayesian inference interpretations), the only property of a machine that matters is which states can be reached from which other states (in a single step) under a given input. (Strictly speaking this only makes sense in the discrete case, but we expect an analogous statement to this to hold more generally.)

This has the consequence that some machines only admit trivial interpretations. By a trivial interpretation we mean one where the posterior is always equal to the prior. Such interpretations exist for every machine, because we can always take the model to be such that H does not change over time and S does not depend on H, so that the input $s \in S$ never gives any information about H. That is, in string diagrams, for any machine we can set



(48)

and



(49)

for any choice of distributions q and r. Then the conditions of definition 2 will always be satisfied. This may be shown using string diagram manipulations and the Markov category axioms.

We now show that there is a class of machines that only admit trivial interpreatations. Consider a machine with the property that $\gamma(y' \,|\, y, s) > 0$ for all y, y', s. That is, for a given input s and initial state y, every state y' has some nonzero probability of being the next state. In this case, eq. (45) implies that

$$\psi_{S,H'}(h, s \,|\, y) = \psi_S(s \,|\, y)\psi_H(h \,|\, y'), \tag{50}$$

for all $y, y' \in Y, s \in S, h \in H$. Since the left-hand side does not depend on y' it follows that $\psi_H(h \,|\, y')$ must not depend on y' either. That is, $\psi_H(h \,|\, y') = p(h)$ for some fixed distribution p. (The other possibility would be that $\psi_{S,H'}(h, s \,|\, y) = \psi_S(s \,|\, y) = 0$, but this can't hold for all $s \in S$, because $\psi_{S,H'}(h, s \,|\, y)$ must be nonzero for some s in order to be normalised.)

This means that if a machine satisfies this property then the only intrepretations it admits are trivial ones, of the form eq. (49). This means that in order for a discrete machine to admit *any* non-trivial Bayesian filtering interpretation it must satisfy a fairly strong constraint, namely that some of its transition probabilities are zero.

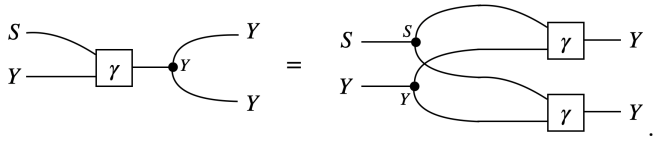
This is to some extend a consequence of our choice to consider only exact Bayesian filtering interpretations. If a discrete machine has no non-zero transition probabilities it might still be possible to interpret it as performing some form of approximate inference, but defining such interpretations precisely is a task for future work.

B.2 More on Bayesian filtering

In this section we show that definition 2 does indeed correspond to Bayesian filtering, at least in the case of a deterministic machine. Our proof of this is inspired by [24, theorem 6.3], which proves an analogous fact about conjugate priors. The proof we give uses string diagram reasoning, which means that it holds even in the most general measure-theoretic context; we do not need to assume that the sets involved are discrete.

Since we restrict ourselves to only deterministic machines in this section, we will note a couple of things about deterministic machines before we talk about Bayesian filtering.

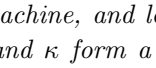
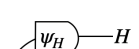
We first note that the condition for a machine γ to be deterministic is



(51)

This comes from the defining equation for deterministic morphisms, eq. (37), and also the axiom (35), noting that γ is a kernel with input $S \otimes Y$ and output Y.

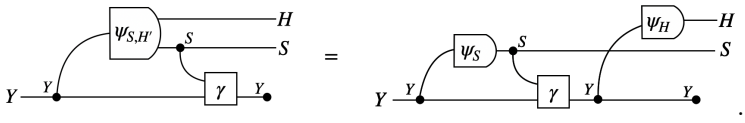
Next we prove the following proposition, which is useful for reasoning about Bayesian filtering interpretations of deterministic machines.

Proposition 1. *Suppose*  *is a deterministic machine, and let*  *and*  *be arbitrary Markov kernels. Then ψ_H and κ form a consistent Bayesian filtering interpretation of γ (i.e. definition 2 is satisfied) if and only if*

(52)

with $\psi_{S,H'}$ and ψ_S as defined in eqs. (3) *and* (4)*.*

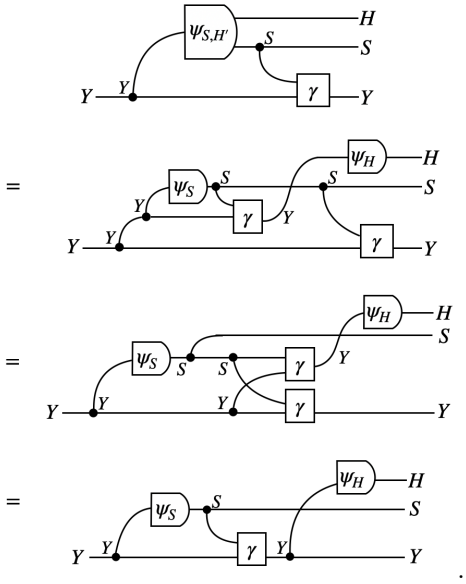
Proof. To see that definition 2 implies eq. (52) we marginalise eq. (5):



(53)

This implies eq. (52) by the rules for Markov categories, specifically eqs. (28) and (32).

For the other direction we assume eq. (52) holds and calculate

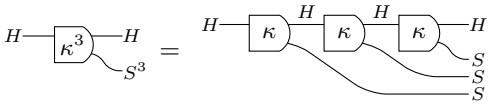


(54)

The first step substitutes in the right-hand side of eq. (52), the second rearranges using the rules of Markov categories, and the third uses the determinism condition. This proves that eq. (5) holds.

We now consider what a Bayesian filtering task involves. The idea is that the reasoner has a model of a hidden Markov process, given by the kernel  . As described in the main text, this kernel can be thought of as a process that simultaneously transforms the hidden state, stochastically, into a new value and emits a visible "sensor value."

Given a kernel of this type, we can iterate it to produce sequences of values in S. For example, we can write

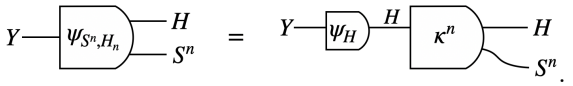


(55)

where S^3 means $S \otimes S \otimes S$ and κ^n is notation for iterating the kernel n times. A kernel of this kind, thought of as an infinitely iterated process, is sometimes called a "coalgebra," since it is a special case of a more general concept of that name. (e.g. [25] takes a coalgebraic approach to de Finetti's theorem.)

For filtering we are interested in inferring the final hidden state of a system, given a finite sequence of visible states. In order to reason about this, we define

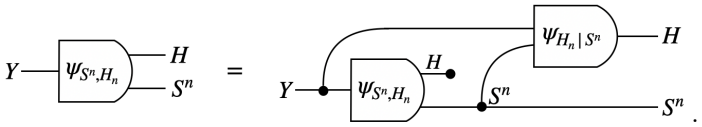
the following kernel:



(56)

This can be seen as an interpretation map, mapping the state of a reasoner to its beliefs about its next n inputs, $S^n = (S_1, \ldots, S_n)$, along with the final value of the hidden state, H_n. These take the form of a joint distribution between S^n and H_n. This joint distribution is formed from the reasoner's initial prior over the initial hidden state H_1 (given by the kernel ψ_H) and the model κ, which is iterated n times.

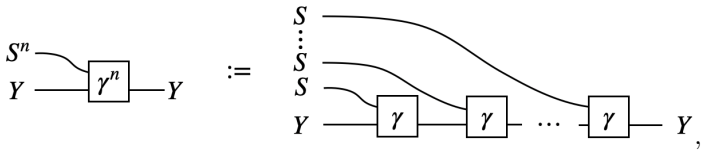
We define this because in filtering we wish to make a probabilistic inference of the final hidden state, H_n, given the sequence of visible states S^n. To infer H_n given S^n we seek a disintegration of ψ_{S^n,H_n}. (See eq. (38) in appendix A.3.) Specifically, we seek a kernel $\psi_{H_n|S^n} \colon S^n \otimes Y \to P(H)$ such that



(57)

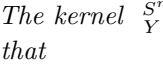
The kernel $\psi_{H_n|S^n}$ takes in a sequence S^n of observations and returns the reasoner's conditional beliefs about H_n, given the sequence S_n. It is also a function of the reasoner's initial beliefs $y \in Y$.

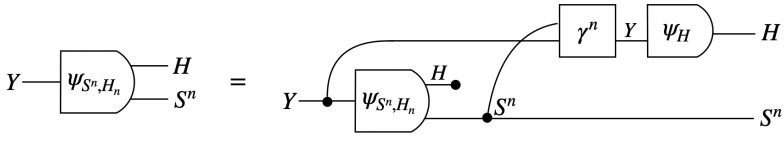
In fact such a kernel can be constructed iteratively in a natural way, if we assume that ψ_H and κ form a consistent Bayesian filtering interpretation. To do this, we first define the iteration of γ, in a similar way to the iteration of κ:



(58)

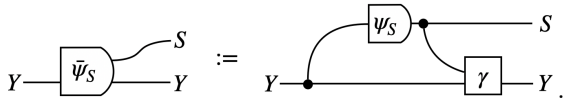
where there are n copies of γ on the right-hand side. We can then state the following result, which shows that consistent Bayesian filtering interpretations can indeed be seen as performing Bayesian filtering, in the discrete case.

Proposition 2. *The kernel*  *is a conditional of* ψ_{S^n,H_n}*, satisfying eq.* (57)*, in that*



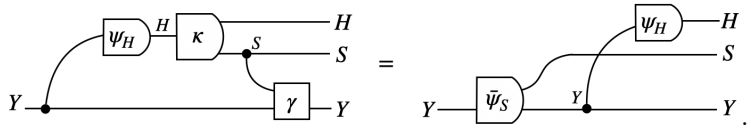
(59)

Proof. We begin by defining the kernel



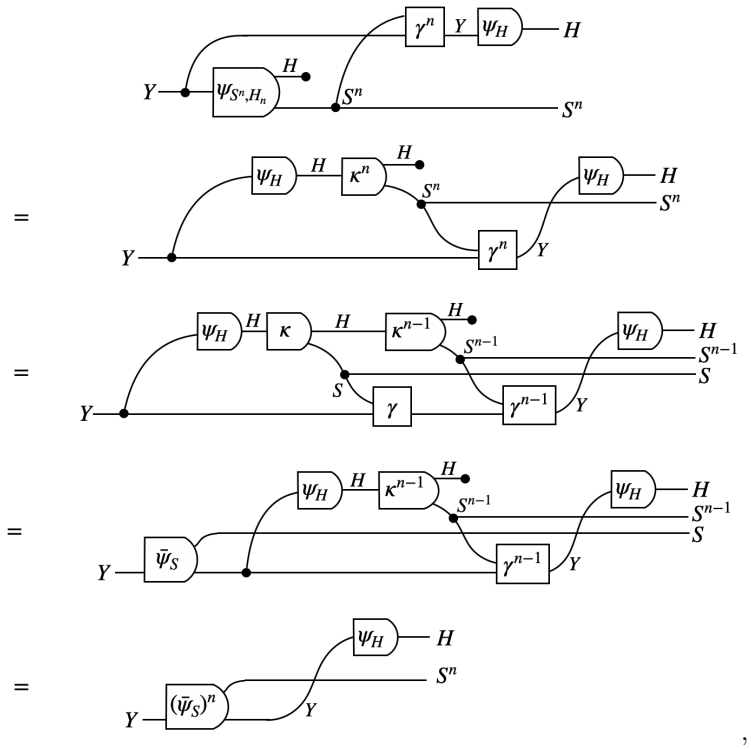
(60)

We also define its iteration, $(\bar{\psi}_S)^n : Y \to Y \otimes S^n$, analogously to κ^n and γ^n. We note that the consistency equation for Bayesian filtering interpretations, eq. (5), can be written in terms of κ and $\bar{\psi}_S$, as



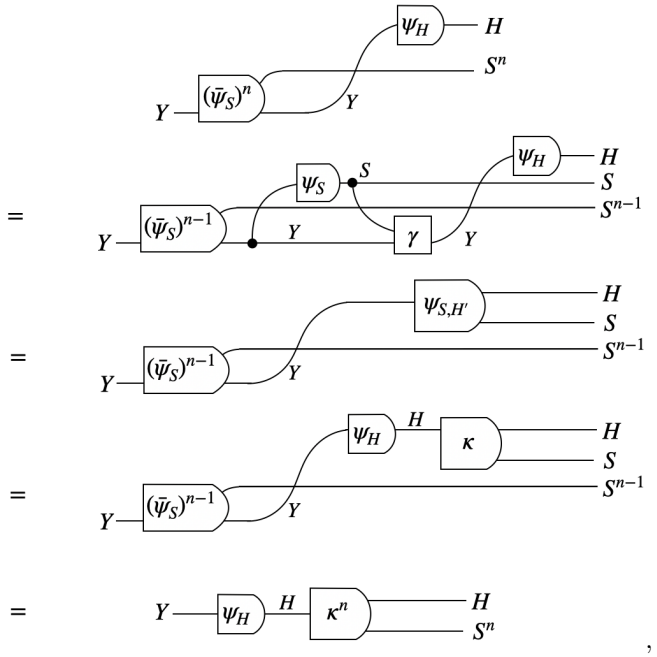
(61)

We then calculate



(62)

where the last step is by applying the other steps inductively. We can then apply a second inductive argument in "the other direction" using eq. (52), as follows:



(63)

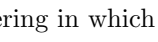
where the last step is again by applying the other steps inductively.

We have proved that $\overset{S^n}{Y}\!\!-\!\boxed{\gamma^n}\!-\!Y\!-\!\boxed{\psi_H}\!-\!H$ is a conditional of ψ_{S^n,H_n}. The kernel $\overset{S^n}{Y}\!\!-\!\boxed{\gamma^n}\!-\!Y\!-\!\boxed{\psi_H}\!-\!H$ can be thought of as giving the reasoner's beliefs about H after receiving a given sequence S^n of inputs, starting from a given initial state $y \in Y$. The result shows that these beliefs are consistent with the agent's prior $\psi_H(y)$ and the model κ, in the sense that the agent's final posterior beliefs about H are a conditional of its initial joint beliefs about the sequence S^n and the final hidden state. We conclude that a deterministic machine with a consistent Bayesian filtering interpretation can indeed be seen as performing a Bayesian filtering task. We expect this to be true in the general case of stochastic machines as well.

B.3 Bayesian inference interpretations and conjugate priors

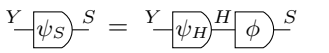
In the main text we noted that Bayesian inference corresponds to a special case of Bayesian filtering. By "Bayesian inference" here we mean the case where the reasoner is interpreted as assuming its inputs are i.i.d. samples from some known distribution with an unknown parameter space H, which we also call the hypothesis space.

The difference between inference and filtering is that we interpret the reasoner as believing that the value of H is unknown but fixed. That is, the reasoner

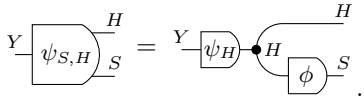
assumes that H doesn't change over time. This corresponds to a special case of filtering in which  , for some kernel ϕ that we also call the model.

While κ can be seen as a model of the environment's dynamics, ϕ has more of the character of a statistical model. It is a model of how the agent's sensor values depend on the unknown value of the hidden parameter H. However, we do not put any constraints on the hypothesis space H or the model ϕ. In particular, we do not assume that ϕ is an injective function $H \to P(S)$, and we allow the case where H is a finite set.

In the case of inference rather than filtering, the kernels ψ_S and $\psi_{S,H'}$ from eqs. (3) and (4) can be written

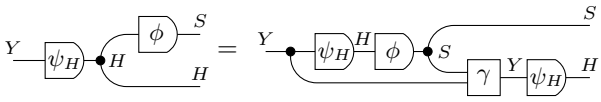

(64)

and


(65)

We write $\psi_{S,H}$ instead of $\psi_{S,H'}$ because in the i.i.d. inference case there is only one hidden variable, that is, $H' = H$. Thus, the joint distribution $\psi_{S,H}(y)$ can be seen as the reasoner's joint belief about its next input and the hidden variable H, when its underlying machine is in state y. The consistency equation for Bayesian inference, eq. (6), then follows by substituting these for $\psi_{S,H'}$ and ψ_S in eq. (5), the consistency equation for Bayesian filtering interpretations.

As with Bayesian filtering interpretations, it is useful to consider the case in which the underlying machine is deterministic (but not necessarily discrete). In proposition 1 we gave a simpler version of the consistency equation for Bayesian filtering interpretations, which is equivalent to definition 2 in the case of a deterministic machine. In the inference case we can substitute eqs. (64) and (65) into this simplified consistency equation (eq. (52)) to obtain


(66)

This is exactly the equation given by [24, eq. 16] as a definition of a conjugate prior.

Both sides of eq. (66) express a joint distribution between S and H, as a function of Y. In the context of conjugate priors, ϕ is considered to be a family of distributions, with parameters H. Our interpretation map ψ_H corresponds to another family of distributions, which is a conjugate prior to ϕ. The machine state Y corresponds to the so-called hyperparameters, i.e. the parameters of ψ_H.

This shift in perspective makes sense. In a computational context, conjugate priors are often useful precisely because they offer a way to perform inference without needing to directly calculate Bayesian inverses at run-time. Instead, the implementation only needs to keep track of the hyperparameters and update

them in response to data. This update takes place according to a deterministic function, whose form depends on the family ϕ and its conjugate prior ψ_H. This updating of the hyperparameters is the role played by γ: it takes in a data point in S along with the current value of the hyperparameters, and returns the updated hyperparameters. Equation (66) asserts that this must done in such a way that the new value of Y does indeed correspond to the correct Bayesian posterior, when mapped to a distribution over H by the kernel ψ_H.

We note that it is somewhat nontrivial to find a pair of kernels ψ_H, ϕ and a function γ such that eq. (66) is obeyed. However, many such examples are known. (Although it is not an authoritative source, a useful list can be found online [40, under "Table of conjugate distributions"], which explicitly gives both kernels and the update function for each example.) Any example of a conjugate prior can be seen as a deterministic machine together with a consistent Bayesian inference interpretation. In addition, in appendix C we give a number of examples of a different flavour, in that in our examples H is either a finite or a countable set.

B.4 Unpacking Bayesian inference interpretations

We now unpack definition 3 by converting eq. (6) into more familiar terms in the case where all the spaces are discrete sets, as we did for filtering interpretations in appendix B.1.

In the case where Y, H and S are finite sets, eq. (6) can be written as

$$\psi_H(h \,|\, y)\, \phi(s \,|\, h)\, \gamma(y' \,|\, s, m) = \psi_S(s \,|\, y)\, \gamma(y' \,|\, s, y)\, \psi_H(h \,|\, y'), \tag{67}$$

or equivalently,

$$\gamma(y' \,|\, s, y) > 0 \quad \Longrightarrow \quad \psi_H(h \,|\, y)\, \phi(s \,|\, h) = \psi_S(s \,|\, y)\, \psi_H(h \,|\, y'), \tag{68}$$

since we can cancel $\gamma(y' \,|\, s, y)$ if we assume it is positive. For $\gamma(y' \,|\, s, y)$ to be positive means that it is possible for the machine to transition from state $y \in Y$ to state $y' \in Y$ after receiving the input $s \in S$.

We can now give an intuitive interpretation to the terms in this equation. If the machine starts in state y, receives input s, and transitions to state y' as a result, then we can regard $\psi_H(h \,|\, y)$ as the reasoner's prior beliefs about the hypothesis h, $\psi_S(s \,|\, y)$ as its prior beliefs about the input s, and $\psi_H(h \,|\, y')$ as the reasoner's posterior belief about the hypothesis h. Equation (68) can then be compared, term by term, to the much more familiar equation

$$p(h)\, p(s \mid h) = p(s)\, p(h \mid s). \tag{69}$$

Here we have written $p(s \mid h)$ in place of $\phi(s \,|\, h)$ and $p(h \mid s)$ in place of $\psi_H(h \,|\, y')$ in order to emphasise the similarity to Bayes' theorem in a more familiar form. Our definition, in the form of eq. (6) or eq. (67), differs from this in that it explicitly takes account of the machine's state, and ϕ and ψ_H are defined by Markov kernels rather than conditional probabilities.

We note that, as in the case of filtering (appendix B.1), our definition of a consistent Bayesian inference interpretation allows the posterior to be arbitrary in the case of subjectively impossible inputs, i.e. those $s \in S$ for which $\sum_{h \in H} \psi_H(h \,|\, y)\phi(s \,|\, h) = 0$ for a given state $y \in Y$. Given such an input the reasoner may update its posterior to anything at all. As with filtering, we regard this as the minimal assumption we could have made, but we can imagine several other choices that one could make instead. These include allowing the posterior to be undefined in such cases; *requiring* it to be undefined; requiring it to obey some additional consistency equation such that the posterior would make sense even on subjectively impossible inputs; or requiring ϕ and ψ_H to be such that subjectively impossible inputs do not exist. We would consider these to be subtly different kinds of interpretation, and we leave their further investigation to future work.

C Details of examples

C.1 An Interpretation Of A Non-Deterministic Finite Machine

We here present a non-deterministic finite machine with internal state space $Y = \{y_0, y_1, y_2\}$ and sensory input space $S = \{s_1, s_2\}$. One Bayesian interpretation

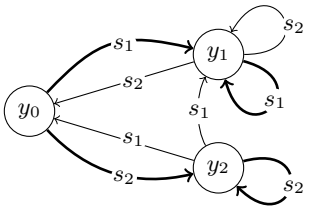


Fig. 1. Transitions for a three-state machine. Deterministic transitions are depicted using bold arrows, and non-deterministic transitions using regular arrows. The precise probability values for non-deterministic transitions are not shown, since we only need to know that they are non-zero.

of this machine, for a hidden state space $H = \{h_1, h_2\}$, is as follows (where δ is the Kronecker delta):

$$\begin{aligned} \phi(s_i \,|\, h_j) &= \delta_{ij} \\ \psi(h_i \,|\, y_j) &= \begin{cases} \delta_{ij} \text{ if } j \in \{1, 2\} \\ 0.5 \text{ if } j = 0. \end{cases} \end{aligned} \tag{70}$$

Under this interpretation, the model ϕ ascribed to the machine is that sensory inputs transparently reflect the hidden state. The machine, in internal state y_0, is taken to be uncertain about the hidden state; in state $y_i \in \{y_1, y_2\}$ it is taken to be certain that the hidden state is h_i. The dynamics of the machine match this interpretation: it transitions deterministically to y_i when receiving input s_i, unless s_i is "subjectively impossible" (s_2 at m_1, and s_1 at m_2). Behaviour on subjectively impossible inputs is not constrained by the consistency equation, so this is a consistent Bayesian interpretation.

C.2 Machine counting occurrences of different observations

We now consider a countably infinite deterministic machine (Y_0, S, γ_0). Let $Y_0 = \mathbb{N}^+ \times \mathbb{N}^+$ ($\mathbb{N}^+$ excludes 0) and the input space be $S = \{+1, -1\}$. The machine deterministically computes the function $f_0 : (Y_0 \times S) \to Y_0$, in the sense that $\gamma_0(f_0(y, s) \,|\, y, s) = 1$. Essentially, it keeps distinct count of how many $+1$ and -1 inputs it has received. Formally:

$$f_0((i, j), s) = \begin{cases} (i+1, j) \text{ if } s = +1 \\ (i, j+1) \text{ if } s = -1 \end{cases} \tag{71}$$

One consistent Bayesian interpretation (ψ_0, ϕ_0) for machine γ_0 uses hypothesis space $H_0 = [0, 1]$ and model:

$$\phi_0(s \,|\, h) = h^{\delta_{-1}(s)}(1-h)^{\delta_{+1}(s)} \tag{72}$$

where

$$\delta_u(v) := \begin{cases} 1 \text{ if } u = v \\ 0 \text{ else.} \end{cases} \tag{73}$$

This model is known as the categorical distribution for two outcomes (or just the Bernoulli distribution). The machine states were deliberately chosen to be the hyperparameters of a possible interpretation map $\psi_0 : Y_0 \to H_0$ which is known as the Dirichlet distribution (and in this special case also as Beta-distribution):

$$\psi_0(h \,|\, (i, j)) = \frac{1}{B(i, j)} h^{i-1}(1-h)^{j-1} \tag{74}$$

where $B(i, j)$ is the Beta function.

This interpretation map (the Dirichlet distribution) is the conjugate priors for categorical distributions. This implies that (ψ_0, ϕ) form a consistent Bayesian inference interpretation, as explained in appendix B.3.

C.3 Machine tracking differences between the number of occurrences of different observations

We now consider another countably infinite deterministic machine (Y_1, S, γ_1) which has the same input space as the machine in appendix C.2. Let $Y_1 = \mathbb{Z}$

and the input space again be $S = \{+1, -1\}$. The machine γ_0 deterministically computes a function $f_1 : (Y_1 \times S) \to Y_1$, in the sense that $\gamma_1(f_1(y, s) \,|\, y, s) = 1$. Machine γ_1 only counts how many more $+1$ inputs it has received than -1 inputs. Formally:

$$f_1(k, s) = k + s. \tag{75}$$

One consistent Bayesian interpretation (ψ_1, ϕ_1) for machine γ_1 uses hypothesis space $H_1 = \{h_{+1}, h_{-1}\}$ and model:

$$\phi(s \,|\, h_i) = \begin{cases} 0.75 \text{ if } i = s \\ 0.25 \text{ otherwise.} \end{cases} \tag{76}$$

The interpretation map $\psi_1 : Y_1 \to H_1$ is

$$\begin{aligned} \psi_1(h_{+1} \,|\, k) &= \frac{1}{2(1 + 0.75^k 0.25^{-k})} \\ \psi_1(h_{-1} \,|\, k) &= \frac{1}{2(1 + 0.75^{-k} 0.25^k)} \end{aligned} \tag{77}$$

It is relatively easy to verify that (ψ_1, ϕ_1) is a consistent Bayesian interpretation with γ_1's dynamics.

As a teaser for future work we may note the following. Since machine γ_0 of appendix C.2 stores the individual counts for s_0 and s_1 inputs, it also implicitly keeps track of the difference between those counts; γ_1 only keeps track of this difference. Consequently, we can define a deterministic kernel $g : Y_0 \to P(Y_1)$ that maps any state $(i, j) \in Y_0$ of γ_0 to $g(i, j) \coloneqq \delta_{i-j}$ which is a probability measure over the state space of γ_1. It turns out that for this map, for any $k' \in Y_1, s \in S$ and $(i, j) \in Y_0$ we have

$$(\gamma_0 \fatsemi g)(k' \,|\, (i, j), s) = \sum_k \gamma_1(k' \,|\, k, s) g(k \,|\, (i, j)). \tag{78}$$

This implies that we can construct an interpretation of machine γ_0 from the interpretation (ψ_1, ϕ_1) of γ_1. For this we precompose the interpretation map ψ_1 for γ_1 with the machine map g to get a consistent Bayesian inference interpretation $(g \fatsemi \psi_1, \phi_1)$ for γ_0. In future work we intend to further develop the theory of how a consistent interpretation of one deterministic machine can be "pulled back" to other machines that are related in a similar way to eq. (78).

D Details on the relation to the FEP

We here try to identify the structures in the FEP that are analogous to the notions of machine γ, model κ, and interpretation map ψ_H. This suggests that, at least in some treatments of FEP, there is an implicit concept that is close to what we have called a reasoner. We will call this putative concept the FEP reasoner.

Large parts of the FEP literature do not explicitly deal with FEP reasoners but are sometimes presented as based on them (e.g. in [20]). The parts that construct the FEP reasoner are those called "Bayesian mechanics" and are still evolving. A standard reference is [19] but this is known to contain some issues [9,21,1]. The most recent version can be found in [16].

Understanding more precisely the relationship between the concepts of our Bayesian and the FEP reasoner is future work. The following are preliminary observations.

D.1 Machine

We first identify the structure in the FEP setup that is most closely related to a machine and is also said to appear to perform Bayesian inference. Unfortunately, the latest iteration of the conditions under which there exists an FEP reasoner, which is [16], does not make this particular structure as explicit as the previous version [19]. We will therefore identify this structure in the older version. A corresponding structure should also exist in the newer version and we will hint at how it may differ.

The FEP setup in [19] consists of four sets of variables $\eta \in E, s \in S, a \in A, \mu \in M$ called external, sensory, active, and internal states with E, S, A, M finite dimensional real vector spaces. These variables obey the stochastic differential equations

$$\begin{aligned} \dot{\eta} &= f_\eta(\eta, s, a) + \omega_\eta \\ \dot{s} &= f_s(\eta, s, a) + \omega_s \\ \dot{a} &= f_a(s, a, \mu) + \omega_a \\ \dot{\mu} &= f_\mu(s, a, \mu) + \omega_\mu \end{aligned} \tag{79}$$

where $\omega_\eta, \omega_s, \omega_a, \omega_\mu$ are independent Gaussian noise terms. The FEP goes beyond the scope of a reasoner and formulates a concept of agent. The concept of an agent should, as part of its interpretation, make it possible to talk about deliberate actions. In the FEP deliberate actions are associated to the active states a. At the same time, the internal states are only involved in inference (or filtering) and the special case where there are no active states seems to be within the scope of the FEP. This should still leave us with a FEP reasoner and make it more comparable to our Bayesian reasoner. We therefore consider the special case where there are no active states such that we get:

$$\begin{aligned} \dot{\eta} &= f_\eta(\eta, s) + \omega_\eta \\ \dot{s} &= f_s(\eta, s) + \omega_s \\ \dot{\mu} &= f_\mu(s, \mu) + \omega_\mu. \end{aligned} \tag{80}$$

This looks like a continuous time version of the Bayesian network in eq. (1) and has the somewhat significant feature that all influences from the external states η are mediated by the sensory states s. This suggests that it is possible to see the sensory states $s \in S$ as inputs to a machine state $\mu \in M$ with the external states $\eta \in E$ "hidden behind" the sensory states.

The internal states $\mu \in M$ are supposed to appear to infer the external states. So the state space Y of the machine of the FEP reasoner should be identified with M. Going by their name and their role in the earlier dynamics of eq. (79) it seems reasonable to identify the sensory state space S with the input state space (also S in our notation) of the machine.

This brings us to the machine's kernel γ. Our formalism does not deal with continuous-time kernels at the moment so we only make some informal comments here. Note that none of the following statements should be considered as proven. Since all variables together form a (time-homogeneous) Markov process, we can choose times $t, t+\tau$ with $\tau > 0$ and write the conditional probability density (assuming things are well behaved enough) at a state (η', s', μ') at $t+\tau$ given a state (η, s, μ) at time t as $p(\eta', s', \mu', t+\tau \mid \eta, s, \mu, t)$ (this notation is taken from [35, p.31]). We can then marginalise out η' and s' to get $p(\mu', t+\tau \mid \eta, s, \mu, t)$ which looks a bit closer to a machine kernel but still depends also on η. We cannot just drop η from this expression even if we assume eq. (80) holds since within a time interval $[t, t+\tau]$ with $\tau > 0$ the influence from η would propagate through the intermediate values of the sensory states to μ'. Instead we here condition on all those intermediate values of the sensory state between t and $t+\tau$. Write $s[t : t+\tau]$ for a part of the trajectory of the sensory state between t and $t+\tau$ that starts in s. Then, assuming eq. (80) we should get:

$$p(\mu', t+\tau \mid \eta, s[t : t+\tau], \mu, t) = p(\mu', t+\tau \mid s[t : t+\tau], \mu, t). \tag{81}$$

In order to make this look even more like a kernel we may take the limit as $\tau \to 0$ and so we write

$$\gamma(\mu' \mid \mu, s) := \lim_{\tau \to 0} p(\mu', t+\tau \mid s[t : t+\tau], \mu, t) \tag{82}$$

which is just a notation for an expression that hopefully provides sufficient intuition for our purposes.

What is important is that within the system eq. (80) there should be a (continuous-time) machine describing the dynamics of the internal states in response to sensory states.

In [16] the structure of eq. (79) and thus eq. (80) is not stated explicitly. However, the sensory (and usually the active states) are still special due to an additional assumption which is also made in [19,34]. The larger process has to have a stationary distribution $p(\eta, s, a, \mu)$ that factorises according to

$$p(\eta, s, a, \mu) = p(\eta|s, a)p(\mu|s, a)p(s, a) \tag{83}$$

which is referred to as a Markov blanket. With this assumption only, one can no longer assume that the sensory states $s[t : t+\tau]$ can "shield" those states from direct influence by external states, which makes it more difficult to compare the dynamics to our setup. A solution may be to use a continuous-time version of the approach in [36]. Below we ignore this issue and assume that we have the structure of eq. (80).

D.2 Model

For a reasoner we also need a model and an interpretation map. As already mentioned the FEP assumes that the system in eq. (79) has a stationary distribution $p(\eta, s, \mu)$. One purpose of this assumption seems to be the definition of what we call the model. In the language of the FEP literature the stationary distribution defines the generative model. Here, generative model refers to a joint probability distribution over causes (parameters/hidden variables) and observed variables. In [16, Section 3.b] the generative model is defined to be $p(\eta, s, \mu)$ with η as the hidden variables and observed variables (s, μ). This could mean that the machine state μ itself is also modelled by an FEP reasoner, which is different from our framework. This would need further investigation that we leave for future work. So we resort to a previous version where only the marginalised stationary distribution $p(\eta, s)$ was considered as the generative model ([34, Fig.3],[19, p.101]). In that case the hidden variable space H in our notation should be identified with the external state space E and the model (in our sense) is a conditional distribution induced by the stationary distribution:

$$\phi(s \mid \eta) \coloneqq p(s \mid \eta). \tag{84}$$

Note that, this choice of a model by itself does not immediately tell us whether the FEP reasoner does filtering or just inference in the sense of definition 3. A model like $\phi(s \,|\, \eta)$ can be part of a filtering kernel κ as well. In both cases we also need an interpretation map.

D.3 Interpretation map

For the interpretation map ψ_H we need a kernel of type $M \to P(E)$. Indeed, a kernel that has the right type can be identified in the FEP literature. This kernel is denoted $q_\mu(\eta)$ and we will identify $\psi_H(\eta \,|\, \mu) = q_\mu(\eta)$. The kernel's definition, however, relies on another assumption of the FEP, namely the existence of a "synchronisation map" $\sigma : M \to E$. To construct σ let us first define two other functions $g_M : S \to M$ and $g_E : S \to E$ via

$$\begin{aligned} f_M(s) &\coloneqq \mathbb{E}_{p(\mu|s)}[\mu] \\ f_E(s) &\coloneqq \mathbb{E}_{p(\eta|s)}[\eta] \end{aligned} \tag{85}$$

and then set

$$\sigma(\mu) \coloneqq f_E(f_M^{-1}(\mu)) \tag{86}$$

which is assumed to be well defined. For details on when this exists in the linear case see [1,16]. With this we can define $q_\mu(\eta)$ and in turn the interpretation map ψ_H. This maps an internal state μ to the Gaussian distribution with mean value equal to $\sigma(\mu)$:

$$\psi(\eta \,|\, \mu) \coloneqq q_\mu(\eta) \coloneqq \mathcal{N}(\eta; \sigma(\mu), \Sigma(\mu)) \tag{87}$$

where the variance $\Sigma(\mu)$ is defined as the variance of the best Gaussian approximation to the model $p(s|\eta = \sigma(\mu))$ when the external state is equal to $\sigma(\mu)$ [34, Eq.2.4]. Note that in [16] the whole stationary distribution is assumed as Gaussian and so $p(\eta|\mu)$ in the corresponding equation in that publication (i.e. Eq.3.3) is also a Gaussian.

In conclusion, the necessary ingredients for something like a Bayesian reasoner seem to be present in the FEP literature. One thing that is special about the FEP reasoner is that its model κ and interpretation map ψ_H are derived from features of the process that the machine is embedded in.

We do not know whether there is an appropriate notion of consistency equation that the FEP reasoner obeys. Presumably, instead of the equation for exact inference that we have presented, such an equation would express the idea that the FEP reasoner performs approximate inference in the form of free energy minimisation. Other differences are that the FEP takes place in continuous time, and perhaps more significantly, that it deals with deliberate actions as well as inference. However, it is not inconceivable that these could be expressed in the form of a consistency equation.

In the current formulations of the FEP, the interpretation is derived from the properties of the 'true' environment, such as the stationary distribution, or the synchronisation map σ. In our consistency equation approach, this need not be the case, since a reasoner's beliefs only need to be consistent and need not be correct. This means in particular that no stationarity assumption is needed.

Nonetheless, perhaps an important idea behind the FEP is that the model that most closely corresponds to the true environment can be considered the best one. A consistency equation approach would still be helpful, in order to systematically explore whether and how interpretations should relate to the larger process in which the machine is embedded.

Blankets All The Way Up – The Economics of Active Inference

Morten Henriksen[0000-0002-5903-7639]

Ministry of Defence, Herningvej 30, 7470, Denmark
mhenriksen84@gmail.com
acw-oe-01@mil.dk

Abstract. A direct implication of active inference, by way of minimizing expected free energy, is the ability to reframe optimization problems as they relate to biological systems. Instead of employing objective functions in order to maximizing an agent's exposure to some exogenous measurable quantity, active inference describes how biological systems optimize by minimizing a divergence (KL) between a posterior probability density and a generative density, by definition endogenous to the system. This particular framework can be shown to underwrite many seemingly disparate disciplines in economics, and may prove to be a source of new insights for the field.

Keywords: Active Inference, Complexity Economics, Behavioural Economics, Decision Theory.

1 Introduction

The free energy principle states that any biological organism capable of existing over a period of time must minimize entropy/surprise, formally described as minimizing a bound on free energy, or minimizing a KL divergence between a posterior density and a generative density. This moves the objective away from a maximization scheme of external quantities, towards the minimization of an internal energy bound afforded by a generative model from which external states are inferred. When taking the expectation, the imperative now becomes to optimize beliefs about world states, rather than maximizing the expected utility of a world state [1], [2], [3]. Minimizing expected free energy can therefore be seen as a way for any system to minimize entropy with respect to a policy, or a plan of action, connecting present states with future states [4]. This particular framework can be shown to underwrite many seemingly disparate disciplines in economics, and may prove to be a source of new insights for the field. In having an active inference framework underwriting economics, as opposed to the

more dominant rational expectations-based general equilibrium approach [5], there will naturally be a move away from "traditional" neo-classical economics and comparative statics, towards a greater appreciation of complexity. In general this opens the field of economics up to a wider area of research in which contributions from neurology, psychology, biology ecology, information- and complexity theory can aid in the further understanding and modeling of economic systems. In particular, active inference provides a first principle account from which complex systems can evolve, and hereby a basis for hypothesis and theory generation to areas of inquiry dominated by a more computationally inspired approach, as is the case in the field of complexity economics [6], [7].

2 Expected Free Energy and Active Inference

While the minimization of variational free energy is a general principle governing how organic (self-organizing) systems exist over time, minimizing expected free energy can be interpreted as extending this principle in order to account for planning [8]. It is in this arena that active inference takes form, and likewise the arena in which the free energy principle could prove useful to the field of economics.

If we start by stating the general principle, free energy (F) can be written as:

$$F = D_{KL}[Q(S_t|O_t;\emptyset)\|P(O_t,S_t)]. \quad (1)$$

Where $Q(S_t|O_t;\emptyset)$ is a variational posterior, $Q(S_t)$ is a variational prior and Ø is a variational parameter.

This can be decomposed into:

$$F = E_{Q(S_t|O_t;\emptyset)}\left[ln\frac{Q(S_t|O_t;\emptyset)}{P(O_t,S_t)}\right]$$
$$= -E_{Q(S_t|O_t;\emptyset)}[lnP(O_t|S_t)] + D_{KL}[Q(S_t|O_t;\emptyset)\|P(S_t)]. \quad (2)$$

Resulting in a (negative) accuracy/energy term, plus a complexity/entropy term, specifying the objective to minimize entropy or complexity in order to maximize accuracy. When minimizing expected free energy (G), the objective becomes to minimize the probability distribution of a policy (path integral) $P(\pi)$, such that $G = -\ln P(\pi)$. Here an optimal policy, denoted $Q^*(\pi)$, will be given by $\sigma(\sum_t^T G_t(\pi))$, where $\sigma(x)$ is a softmax function. We can therefore write:

$$G_t(\pi) = E_{Q(O_t,S_t|\pi)}[ln\ Q(S_t|\pi) - ln\ \tilde{p}(O_t|S_t)]$$
$$\approx E_{Q(O_t,S_t|\pi)}[ln\ Q(S_t|\pi) - ln\ \tilde{p}(O_t) - lnQ(S_t|O_t)]$$
$$\approx -E_{Q(O_t,S_t|\pi)}[ln\tilde{p}(O_t)] - E_{Q(O_t|\pi)}D_{KL}[Q(S_t|O_t)\|Q(S_t|\pi)]. \quad (3)$$

[9].

Decomposing expected free energy into an extrinsic value term, often described as a goal directed term or pragmatic value, and an intrinsic value term, which can be interpreted as an epistemic value term, or simply information gain [10]. Here, minimizing expected free energy becomes a mixture of pragmatic and epistemic considerations where maximizing exposure to information is used to indicate the appropriateness of a policy (epistemic value), given a specific goal (pragmatic value). Naturally there exists a trade-off between epistemic and pragmatic value, where both can be shown in isolation to be equally valid strategies for minimizing expected free energy, but in different ways. However, when connecting future states with present states by selecting a trajectory expected to minimize surprise, this trajectory can only be propositional when considering complexity. The difference is comparable to planning and executing a move from A to B considering what is known about the environment, or planning and executing a move from A to B considering what is known about the environment, while in heavy traffic. None withstanding, the principle of least action, or indeed, least effort [11], must apply, given that all optimal policies are referencing a path integral with a minimum expected time average.

Note that the only way for an agent to minimize expected free energy, is to actively bring desired future states into existence through action, and while this may seem like a trivial statement, it points to a very interesting implication regarding the concept of planning as inference [12]. In essence there is only action, encoded as an expected sequence to be performed, here represented as intensity (energy divergence) as a function of time. What this means, is that we can treat an expected action sequence and an expected time sequence interchangeably, the expected time sequence being a function of the expected action sequence and vice versa. This quickly moves us into the realm of time perception, and gives us the ability to hypothesise about the various ways time perception can alter in response to both the work of the system in the present, and the expected work to be done in the future [13], [14]. As such, time perception will be

influenced by surprise, as well as the expected time average of surprise (entropy). If nothing more, this gives us the ability to speak in terms of urgency when considering various policies, and therefore various degrees of "intensity" connected to different utilities. More generally, we observe a system that changes dynamically in response to information, and it is this "adaptive" ability that interferes with more linear formulations of agent behaviour, exemplified in the expected utility theory for instance [3], [5], [6], [7]. This does not mean however, that it becomes impossible to derive general statements about agent behaviour from an active inference perspective, quite the contrary. The presence of a generative model from which external (hidden) states are initially inferred, and the presence of the epistemic value term (information gain) to which the generative model must adapt, actually provides many opportunities for a priori statements concerning agent behaviour and economics more generally.

3 Discounted Utility

Equation 3 can be read as describing the probability of occupying an expected future state given a specific trajectory (policy) generated by the generative model (prior and likelihood). Here there will be an attraction to future states of high expected probability, reflecting what could be interpreted as expected utility, if the prior density is associated with "prior preferences" [15]. The trajectory is controlled by information, where low probability states are equivalent to states of high surprise. As such, low probability states command a low expected utility, the avoidance of which is dependent upon the minimization of surprise or negative log model evidence ($-\ln p(O_t)$). Given the time subscript (t), longer policy sequences will be associated with higher probabilities of occupying more surprising states, reflecting an increase in total entropy over time. This is in part due to the use of approximate Bayesian inference that is a consequence of the variational treatment on free energy show in equation 1. If expected utility then is associated with states of high probability, expected utility must fall over time in proportion to an increase in entropy; naturally giving us discounted utility functions, as well as a general theory of time preference[1].

[1] The observation that goods or services are preferred sooner rather than later, all else being equal. In economics, the prefixes "high" or "low" is sometimes used in order to differentiate

4 Probability and Utility Spaces

Things will, however, become a lot more interesting when considering the full impact of epistemic value or information gain. Here information gain will inherently be uncertainty resolving given the hidden states (S_t) in the environment. Because of this, the objective function governing the policy trajectory cannot take any specific value, or indeed be defined with a terminal, but must in a sense be "discovered" due to the unknown cost functions governed by surprise/information gain. As such, the objective function is simply to minimize more or less hidden cost functions, and by this measure minimize surprise or entropy through adaptation. What this means for the specifics of any given utility function, is that it cannot technically be connected to the utility either X, Y or Z, since this quantity cannot be evaluated at time t_0. What is evaluated is a utility space, in conjuncture with a probability space, describing a set of solutions (utilities) the most optimal of which, is "discovered" through action and adaptation. If, however, priors are allowed to perfectly model hidden states, then active inference could be described without cost functions [16]. This would however render the variational treatment moot, as the variational density under the "true" posterior would reflect what for all intents and purposes looks like exact Bayesian inference. While this approach could prove very useful in modelling autonomic responses, it can only take us so far when considering complexity.

The presence of a utility space, as opposed to specific utility given by revealed preferences[2], means that we can start to combine intuitions, ad hoc observations and various formalisms regarding the notions of heuristics, metaheuristics and satisficing behaviour[3] [17]. At t_0, expected utility will be subject to a discount function, but since this function is a consequence of time sensitive uncertainty about state transi-

between various levels of "impatience". High time preference agents will value time at a high rate and display high levels of impatience, while low time preference agents will display low levels of impatience. High or low time preferences are therefor often used as explanations for various levels of propensities to consume or save in an economy.

[2] Revealed preferences is a way to infer an agents utility function by observing past behaviour. As such, agents cannot change their preferences once they have been revealed, since this would change the utility function and herby greatly complicate economic modelling practices. The concept of revealed preferences is tightly linked to the transitivity axiom [18], [19].

[3] The idea, that in situations where optimal solutions do not exist, an agent will search until a solution is deemed to be good enough.

tion, the utility of any single prospect is not merely residing in a probability space, but must itself be probabilistic in nature. This means that any specific preferred future state, and the utility that this state represents, does not exist as such, prior to occupying the preferred state. What utility instead is "reduced" to, is a policy inferring expected surprise in accordance with least action, describing a categorical representation of preferred future states. Subjective value will therefore be governed by the specifics of the policy trajectory, and not the expected utility of any specific prospect. Intuitively we must admit that any preferred future state is colloquially a "figment of our imagination", as the information describing this state cannot be taken fully into account. In fact, the more specific the description of any future state, the higher the probability of this state being purely fictitious and therefore surprising. Maintaining a specific description of a future state would hereby not be a tenable strategy for optimization given the potential amplitude of the loss functions, and as a consequence, the inability to effectively minimize surprise.

5 Active Inference and Biases

Still occupying t_0, the utility space represents options, that at first glance looks like indifference, that is, equalities between various preferred states or utilities. However, the selection of a policy from the generative model, must favour some states in the utility space over others when this policy is acted out. In concert with information update revealing hidden states in the environment as they occur, continuous movement along a policy trajectory will favour fewer and fewer states until only one remains. This means that the ex post policy trajectory was selected for by the environment in conjuncture with the generative model, making the attached utility or subjective value of a given prospect a function of beliefs about how states in the world unfold, the uncertainty these beliefs entail as an increasing function of time, as well as the actual unfolding of events in the world forcing the need for adaptation. Isolating a single specific good from our utility space now allows us see how the attached utility or subjective value of this good changes along the policy trajectory in response to the minimization of surprise, as the good variably moves further away, or closer to, the agent in time and effort. This also allows us to see why a good once obtained tend to

command a higher subjective value than a comparable or identical good not yet obtained, as is the case with the endowment effect [20]. The higher subjective value place on things close to/visible/understandable/not hidden, similarly aids in explaining the status quo bias, as well as various anchoring effects [20], [21]. The anchoring effect is however a curious case, since active inference demonstrates how "value" is ultimately determined *actively* when beliefs are acted out, rather than simply stated. What agents believe some good is "worth" based on, or not based on, anchoring effects, is not a good indicator of what agents will do in order to obtain a good in the final analysis when cost functions or surprise is taken into account. Interestingly, many of the implications of active inference in economics can be recapitulated in terms of transaction cost economics, where the goal is to formalize and take into account the costs associated with being an economic agent given such things as bounded rationality and imperfect information [22], [23], [24]. The connection between transaction cost economics and active inference, will however not be treated in this presentation, but could prove to be a very interesting area of study going forth.

6 A Simple Model

Ultimately, subjective value is determined by effort, which is to say that effort likewise commands a "value". One aspect is the rate of remuneration[4] (RR) that active inference implies; another is the increasing present value[5] (PV) of various prospects as a function of time preference. The RR can be shown as an increasing function of

[4] There is at present no corollary in the economic literature to a rate of remuneration as used in this presentation. Normally a remuneration rate simply refers to a salary or stream of payments due for work or services rendered. Here, a rate of remuneration refers the expected energy input for a system given energy output, where the minimum requirement is long run homeostasis. For this reason the term may be ill conceived. Conversely, the term as used herein perfectly captures the observation, that agents have a preference for increasing over declining sequences not strictly "permissible" under a rational expectations framework [25], [26].

[5] Normally present value refers to the value at present of a discounted future cash flow where $PV = \frac{CF}{(1+r)^n}$. Here, the term refers to subjective value given a time component. As such, it is the expected value of something that by necessity must lie in the future, and therefore must be discounted to some degree, considering a generative model that takes surprise or uncertainty into account. We can therefore also treat present value and utility (discounted) as interchangeable.

time, while the PV, or inverse discount rate, is a decreasing function of time. As a simplification, or heuristic if you will, we can view the PV curve as reflecting the extrinsic value term in equation 3, and the RR as reflecting the intrinsic value term. To qualify this statement, we can refer to the previous discussion on discount rates concerning extrinsic value. Depicting intrinsic value as an increasing rate of remuneration will however need to be elaborated. Intrinsic value is in the literature [10] often interpreted as information gain when viewing free energy minimization from an information theoretic standpoint. Moving in time reveals hidden states, and to the extent that these states do not match expectations, they will be surprising for the system. Minimizing surprise prompts the system to adapt, generating cost functions (complexity cost) to which a given future prospect must be able to remunerate. Over a longer and longer time interval, the cost functions add up in proportion to a higher and higher rate of remuneration. As such, the RR communicates the "cost of acting" over a time period given a hypothetical discount rate on the future. Overlaying the two functions denoting the X-axis as time/action and the Y-axis as entropy, shows the trade-off between exploration and exploitation, the intersect representing the "sweet spot" between this trade-off. Here the PV (discount) rate will communicate uncertainty or ambiguity with regards to state transitions, and the RR communicates the cost of "foraging" uncertainty resolving information. We can therefore also view the RR as the *active* part of active inference, and the PV as the *inference* part.

While the entropy axis is self-explanatory, the time/action axis can be thought of as the level of urgency in the system, that is, the amount of work to be done in a given time frame. Shifting the PV schedule to the right will hereby depict a system with low uncertainty about state transitions, commensurate with a well explored environment. The intersect is now higher on the RR, where the cost of further exploration likewise is higher, rendering exploitation a relatively more preferable strategy. This is demonstrated by the intersect being further along the time/action axis, commensurate with a less urgent system. Conversely, a leftward shift in PV schedule depicts a system with high uncertainty about state transitions, imitating perhaps a poorly explored environment. Here, the intersect is closer on the time/action axis, as uncertainty forces the system to act in a more urgent manner. The intersect is likewise lower on the RR,

meaning that epistemic foraging comes at a lower cost, rendering exploration a relatively more preferable strategy.

Interestingly, Milling et al. [9] have proposed an amendment to the expected free energy formulation, where information gain is penalized instead of encouraged. Rather than being an alternative approach to expected free energy, one the paper has termed "free energy of the future (FEF)", the two formulations might actually depict two equally valid strategies under different generative models, emphasizing either exploration or exploitation respectively as demonstrated in the model above.

$$\begin{aligned} G_t(\pi) &= E_{Q(O_t,S_t|\pi)}[ln\ Q(S_t|O_t) - ln\ \tilde{p}(O_t|S_t)] \\ &= E_{Q(O_t|\pi)} D_{KL}[Q(S_t|O_t)\|\tilde{p}(O_t|S_t)] \\ &\approx -E_{Q(O_t,S_t|\pi)}[ln\tilde{p}(O_t|S_t)] + E_{Q(O_t|\pi)} D_{KL}[Q(S_t|O_t)\|Q(S_t|\pi)]. \end{aligned} \tag{4}$$

[9].

Where a shift in the PV schedule represents various levels of uncertainty about state transitions, a shift in the RR schedule represents surprise, either positive for a leftward shift or negative for a rightward shift. Here we see how the present value or utility of any given prospect changes in response to surprise, as the prospect variably moves closer to, or further away from the agent in time and effort. The Y-axis, at present denoted simply as entropy, will be analogous to system stability. While the model lacks a dimension necessary to represent this dynamic, it will none the less be an a priori consequence of the underlying mathematics. A low degree of uncertainty about state transitions is commensurate with intersects that lies comparatively higher on the Y-axis, and a high degree of uncertainty about state transitions lies comparatively lower on the Y-axis. In both examples there will be trade-offs, where low uncertainty generates stability, but at the same time, rigidity, locked-in processes and path dependency. High uncertainty generates adaptability and flexibility, as well as volatility and unpredictability commensurate with a lower probability of "correctly" inferring expected state transitions.

7 Conclusion

In this presentation, I hope to have demonstrated how active inference can help to inform the field of economics. Apart from elucidating on various biases in agent be-

haviour, active inference also holds promising implications for price, interest and transaction cost theory by providing a rigid principle for agency and organization. More than this, active inference provides an intuitive and simple avenue for the introduction of complexity into classical or standard economic theory, with the possible implication of being able to aid in the modeling and analysis of economic data and events going forth.

References

1. Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. A tale of two densities: active inference is enactive inference. Adaptive Behavior, 28(4), 225–239 (2020). https://doi.org/10.1177/1059712319862774.
2. Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. . The anatomy of choice: active inference and agency. Front. Hum. Neurosci., 25 September (2013). https://doi.org/10.3389/fnhum.2013.00598.
3. Henriksen, M. Variational Free Energy and Economics Optimizing With Biases and Bounded Rationality. Front. Psychol., 06 November (2020). https://doi.org/10.3389/fpsyg.2020.549187.
4. Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., Pezzulo, G. Active inference and learning. Neuroscience & Biobehavioral Reviews, 68, 862-879, ISSN 0149-7634, 22 June (2016). https://doi.org/10.1016/j.neubiorev.2016.06.022.
5. Arthur, W. B. Foundations of complexity economics. Nat Rev Phys 3, 136-145 (2021). https://doi.org/10.1038/s42254-020-00273-3
6. Arthur, W.B. Complexity and the Economy. Oxford University Press (2015). ISBN 978-0-19-933429-2
7. Farmer, J., Foley, D. The economy needs agent-based modelling. Nature 460, 685-686 (2009). https://doi.org/10.1038/460685a.
8. Parr, T., Friston, K.J. Generalised free energy and active inference. *Biol Cybern* 113, 495–513 (2019). https://doi.org/10.1007/s00422-019-00805-w
9. Milling, B., Tschantz, A. and Buckley, C. L. Whence the Expected Free Energy? Neural Computation, 33 (2), 447-482 (2021), https://doi.org/10.1162/neco_a_01354
10. Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. Active inference and epistemic value, Cognitive Neuroscience, 6:4, 187-214 (2015), DOI: 10.1080/17588928.2015.1020053
11. Kim, C.S. Bayesian mechanics of perceptual inference and motor control in the brain. *Biol Cybern* 115, 87–102 (2021), https://doi.org/10.1007/s00422-021-00859-9
12. Botvinick, M. & Toussaint, M. Planning as Inference. Trends in Cognitive Sciences, 16 (10), 485-488 (2012), https://doi.org/10.1016/j.tics.2012.08.006.
13. Roseboom, W., Fountas, Z., Nikiforou, K. Activity in perceptual classification networks as a basis for human subjective time perception. *Nat Commun* 10, 267 (2019), https://doi.org/10.1038/s41467-018-08194-7

14. Zakharov, A., Crosby, M. and Fountas, Z. Episodic Memory for Learning Subjective-Timescale Models, (2020). arXiv:2010.01430
15. Lopez-Parsem, A., Domenech, P. and Pessiglione, M. How prior preferences determine decision-making frames and biases in the human brain. eLife 2016;5:e20317 (2016), DOI: 10.7554/eLife.20317.
16. Friston, K., Samothrakis, S. & Montague, R. Active inference and agency: optimal control without cost functions. *Biol Cybern* 106, 523–541 (2012). https://doi.org/10.1007/s00422-012-0512-8
17. Simon, H. A. Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138, (1956), https://doi.org/10.1037/h0042769
18. Samuelson, P. A Note on the Pure Theory of Consumers' Behaviour. Econometrica, 5, 61-71 (1938). https://doi.org/10.2307/2548836.
19. Von Neumann, J., Morgenstern, O. Theory of Games and Economic Behaviour (1947) Princeton, NJ: Princeton University Press
20. Kahneman, D., Knetsch, J. L., and Thaler, R. Anomalies: the endowment effect, loss aversion, and status Quo bias. *J. Econ. Perspect.* 5, 193–206, (1991), doi: 10.1257/jep.5.1.193
21. Ariely, D., Loewenstein, G., Prelec, D. "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences, *The Quarterly Journal of Economics*, Volume 118, Issue 1, February 2003, Pages 73-106 (2003) https://doi.org/10.1162/00335530360535153
22. Coase, Ronald H.: The Problem of Social Cost; The Journal of Law and Economics Vol. 3, (1960).
23. Williamson, O. The Economics of Organization – The Transaction Cost Approach; American Journal of Sociology Vol. 87, No. 3, 1981. https://www.researchgate.net/publication/235356934_The_Economics_of_Organization_The_Transaction_Cost_Approach
24. Williamson, O. Transaction Cost Economics - An Introduction; Discussion Paper No. 2007-3, (2007). http://www.economics-ejournal.org/economics/discussionpapers/2007-3
25. Loewenstein, G. F., Prelec, D. Preferences for sequences of outcomes. Psychological Review, 100(1), 91-108, (1993), https://doi.org/10-1037/0033-295X.100.1.91
26. Scholten, M., Read, D. Better is worse, worse is better: Violations of dominance in intertemporal choice. Decisions, 1(3), 215-222, (2014), https://doi.org/10.1037/dec0000014.

Filtered States: Active Inference, Social Media and Mental Health

Ben White[1] and Mark Miller[2]

[1] University of Sussex, Brighton, UK
ben.a.white99@gmail.com
[2] Hokkaido University, Sapporo, Japan
markmiller@chain.hokudai.ac.jp

Abstract. Social media is implicated today in an array of mental health concerns. While worries around social media have become mainstream, little is known about the specific cognitive mechanisms underlying the correlations seen in these studies, or why we find it so hard to stop engaging with these platforms when things obviously begin to deteriorate for us. New advances in computational neuroscience are now perfectly poised to shed light on this matter. In this paper we approach these concerns around social media and mental health issues, including the troubling rise in Snapchat surgeries, depression and addiction, through the lens of the Active Inference Framework (AIF).

Keywords: active inference · social media · depression · addiction.

1 Introduction

Levi Jed Murphy smoulders into the camera. It's a powerful look: piercing eyes, a razor-sharp jawline, and high cheekbones, which according to Levi himself cost around £30,000 pounds [17]. Levi is an influencer from the UK, with a large social media following. Speaking on growing his following, Levi reveals that if a picture doesn't receive a certain number of likes within a set time, then it gets deleted, and that the surgeries are simply a way to achieve this rapid validation: "it's important to be good looking for social media, because obviously I want to attract an audience", he states [47]. While the filter that inspired Murphy's surgeries has since been banned, many similar ones are still available, and Murphy's story highlights growing concerns about a phenomenon now dubbed 'Snapchat surgery'. One survey of young people online found that nearly half had felt influenced by social media to consider cosmetic surgery [4]. While these concerns around social media and cosmetic surgery are recent, they can be added to a litany of worries about the effect of social media use on mental health and general wellbeing.

Today, social media is implicated in an array of mental health concerns. A 2017 report published by a parliamentary group in the UK linked social media use with a range of worries about mental health [9], while a growing number of

empirical studies link social media use with symptoms of addiction and depression [3][33]. Worries that social media platforms might in some way warp our perception of the world, or cause low self-esteem or diminished life satisfaction, seem to be running through the mainstream collective psyche and even some former influencers have begun to turn against various social media platforms, highlighting the dangers of curating a self-image with little purchase on reality [21]. In response, some platforms have begun trialling design tweaks aimed at protecting user's health, such as limiting the visibility of 'likes' on a post.

While concerns around social media have become mainstream, little is known about the specific cognitive mechanisms underlying the correlations seen in these studies, and why we find it so hard to stop engaging with these platforms when things obviously begin to deteriorate for us. In what follows, we suggest that both the rise in Snapchat surgery, and the connections between social media, depression and addiction, can be accounted for via a unified theoretical approach grounded in an emerging, and now highly influential, theory of cognition and affect - the active inference framework (AIF). We propose that the structure of some social media platforms constitute what have been dubbed 'hyperstimulating' digital environments, wherein the design features and functional architecture of digital environments impacts the machinery of cognition in ways which can lead to a warping of healthy agent-environment dynamics, producing precisely the sorts of pathological outcomes we see emerging today [48]. In what follows we will first briefly introduce the AIF. Next, we highlight how the same predictive mechanisms that keep us alive and well can also become warped, leading to aberrant feedback loops in cognition and behavior that help explain the various psychopathologies that are related today to social media and internet use. Finally, we will argue that various digital environments, including social media platforms, have very specific design features and mechanisms that leave our predictive systems particularly vulnerable to these kinds of suboptimal feedback loops.

2 Introducing Active Inference

The revolutionary move of the AIF is to reimagine the brain as a prediction engine constantly attempting to predict the sensory signals it encounters in the world and minimising the discrepancy ('prediction errors') between those predictions and the incoming signal [18][24][7]. To make apt predictions these systems need to build up a 'generative model': a structured understanding of the statistical regularities in our environment which are used to generate predictions. This generative model is essentially a model of our world, including both immediate, task-specific information, as well as longer-term information that constitutes our narrative sense of self. According to this framework, predictive systems can go about minimising prediction errors in two ways: either they update the generative model to reflect the world more accurately, or they behave in ways that bring the world better in line with their prediction [7]. In this way, the brain forms a part of an embodied predictive system which is always striving to move from un-

certainty to certainty. By successfully minimising potentially harmful surprises these systems keep us alive and well. Consider the healthy and highly expected body temperature for a human being of 37°C. A shift in temperature in either direction registers as a spike in prediction error, signalling to the organism that it is moving into an unexpected, and therefore a potentially dangerous, state. If the change in temperature is not too extreme, we could just sit there and come to terms with the changing temperature (update our generative model), or we might reach for a blanket or open a window. In these cases what we're doing is acting upon our environment, sampling the world, and changing our relation to it, in order to bring ourselves back within acceptable bounds of uncertainty.

Predictive systems must be flexible and able to quickly adapt to changing conditions within an environment. According to the AIF the predictive system is flexibly tuned by second order predictions that estimate the salience and reliability of the error units resulting from first order predictions given the current context [39]. So called 'precision weighting' acts to modulate the impact that particular prediction errors have on the system. For example, high precision can drive learning and further processing, while low precision would render a signal relatively impotent within the system [8]. This mechanism allows the system contextual flexibility and can also allow greater reliance on either the generative model or sensory signals. Precision also plays a central role in selecting which behaviours are enacted, as actions are selected based on expectations about future error reduction. That is to say, predictive agents score and select behaviours based on predictions about the likely error-reducing capacities of those behaviors within a given context. In other words, precision is weighted on beliefs about policies, given the likelihood that in a given context a certain policy will lead to a certain reduction in error [19].

Crucial for this process is a sensitivity about how well we are managing error over time relative to expectations. 'Error dynamics' refers to changes in the rate of average error reduction over time [27][10][29]. The rate of change in error can be visualized as a 'slope', with steep decreases in error minimization representing that the system is doing well at confirming predictions, and a steep increase as a loss of predictive acuity. On the agent level, changes in error dynamics are experienced as valenced bodily affect (i.e. positive and negative feelings accompanied by approach or avoidance tendencies). When an organism registers a slope of error reduction in line with (or better than) its expectations, they are 'rewarded' with positively valenced affective changes. When the rate of prediction error rises (or the rate of reduction slows down) the organism is 'punished' with a negatively valenced affect [14]. These affective changes play a role in the AIF by tuning precision weighting on action policies. Positive or negative valence acts as feedback to the system, up-regulating or down-regulating precision expectations respectively. In short, valenced bodily affect shifts us toward a closer attunement with the environment, by raising or lowering the system's confidence in sets of action policies relative to how well or poorly those behaviours have proven themselves to be relative to expectations [30].

This means that as predictive organisms, we actively seek out waves of manageable prediction error - manageable uncertainty - because resolving it results in our feeling good. Predictive organisms that are tuned by error dynamics then will naturally exhibit curiosity and exploratory behaviour [29]. They will be moved affectively to seek out and make the most of the steepest slopes of error reduction in their environment. Situations which offer too little resolvable error (i.e. are too predictable) are boring for such organisms, while situations with too much error (i.e. too uncertain) are experienced by the agent as frustrating or threatening. The recent rise in jigsaw puzzle sales during the covid lockdown testifies to our love of manageable uncertainty. These feelings evolved to keep us well-tuned to our environment, helping us to curiously feel out novel and successful strategies for survival, while also avoiding all the stress and unpleasantness which comes with runaway uncertainty. This active, recursive, and felt relationship with the environment is crucial to grasping how social media can be detrimental to our mental health, and why we often find it so hard to stop using it, as we will see next.

Living well, in active inference terms, means being able to effectively manage uncertainty – and that's predicated on having a generative model which represents the world accurately. A generative model that poorly reflects the regularities of the environment would inevitably lead to an increase in bad predictions, and a flood of difficult-to-resolve errors. Active inference theorists are beginning to develop novel accounts of mental health conditions which focus on the predictive effectiveness of a person's generative model [44][5][34][12]. In the next section we look at how social media threatens to engender these suboptimal generative models in users.

3 Your Brain on Social Media

Social media is a spectacularly effective method for warping our generative models, as it often bombards users with bad evidence about both the world around us and our place in it. Typically, in the offline world, our generative model and expectations are encoded with information incoming from the unfiltered environment, which means that most of the time our generative model accurately (or at least usefully) reflects the world. However, in cases of regular and heavy engagement with social media, incoming information about the world is very often carefully selected, curated, and altered - we're potentially engaging with a fantasy. Moreover, apps that offer the use of filters also allow us to represent ourselves in carefully curated ways, potentially cultivating kinds and quantities of feedback and validation simply not available to us when we go offline. The space between being and appearing is potentially vast - with a few swipes we can dramatically alter our appearance or retake the same picture twenty times until our face exudes the calm mastery of life we want to project. As social media platforms develop features which foster an increasing potential for inauthenticity, the more those platforms become powerful bad-evidence generators, flooding the predictive systems of their users with inaccurate information, telling us that

the world is full of incredibly beautiful, cool people, living wonderfully luxurious lives: social media platforms can act as a digital crowbar, prising apart our generative model from the offline environment. Instead, our model of the real world comes to take on the expectations generated through the online one, and the result is increasingly unmanageable waves of prediction error which the system must now strive to minimise.

The seemingly extreme actions of seeking cosmetic surgery to look more like one's online presence are one strategy for resolving this kind of prediction error. A recent survey found that more than half of cosmetic surgeons had patients ask explicitly for procedures which would enhance their online image, while many also reported patients using enhanced images of themselves as an example of how they'd like to look [26]. Levi the influencer describes how filters allowed him to preview the effects of specific cosmetic procedures, and while Instagram has now banned that specific filter, many perform similar functions. While this may seem extreme, these actions make perfect sense when viewed through the AIF. If we become accustomed to our own doctored appearance, and to receiving all the feedback associated with it, soon the level of validation available offline will be registered as a mounting prediction error, that's likely to result in feelings of stress, and inadequacy. According to the AIF, seeking surgery to bring our offline self in line with our online self is no different from grabbing a blanket as the temperature begins to drop - we're sampling the world to bring us back into an expected state, acting to minimize prediction error. It's just that through very deliberate design features, social media is - for some users - capable of displacing our self-image so much that the only way to rectify the error and meet those expectations is to surgically alter the way we look.

Note, though, how high the stakes are in this scenario. Surgery might offer one way to attempt to resolve the mounting error, but if we're unable to resolve the error, and continue to engage with social media, then this consistent failure is fed back to the system, eventually teaching it to expect its own failure and inability to act effectively in the world. This 'pessimistic' tendency in prediction bears a striking resemblance to the kind of scenarios now described by neuroscientists working on computational accounts of depression based on the AIF. Various forms of psychopathology, including depression, have now been described as a form of 'cognitive rigidity' wherein the system fails to adjust its expectations (including expected rate of error reduction) in line with feedback from the world [30][5][6][40][16][41][46]. In properly functioning predictive systems, when there is failure to resolve error in line with expectations, negatively valenced affect feeds back to the system and downregulates expectations accordingly, which then leads to the system being likely to resolve error once again in line with new expectations, which results in positive affect and an upregulation in expectation [30]. This constant undulation of expectation and valenced affect serves to keep well-functioning agents in a relatively stable state. In AIF accounts of depression however, when error isn't reduced in line with expectations, the system fails to update those expectations, leading to ongoing failure and an inverse slope of error reduction. The long-term summation of error manifests as persistent low mood,

leading to a downregulation of precision on action policies. In short, a system which displays this rigidity in expectation comes to predict its own failure and ineffectiveness, which manifests on the agent level as symptoms of depression, such as feelings of helplessness, isolation, lack of motivation, and an inability to find pleasure in the world [30].

A 2018 exchange between Instagram user 'ScarlettLondon' and Twitter user 'Nathan', illustrates widespread intuitions about a link between social media and depression. 'ScarlettLondon' posted an image of her "morning routine" with a caption reading "I... give you a little insight into how I start my day in a positive way." The image featured Scarlett in a luxurious hotel room, with a selection of breakfast dishes laid out on the bed, complete with a product placement for Listerine. 'Nathan' reposted the image with another caption, reading "Fuck off this is anybody's normal morning. Instagram is a ridiculous lie factory made to make us all feel inadequate" [36]. Nathan's sentiment captures a pervasive intuition, since confirmed in several studies: that social media can cause depression because it facilitates negative comparisons with inauthentic or otherwise unattainable content [11].

Indeed, engagement with social media platforms has been shown to have a measurable impact on an individual's expectations of a specific place or event [37]. Through ongoing and consistent engagement with inauthentic content, a user's expectations for successful error reduction in the environment have the potential to effectively be 'pinned' in place, leading to the predictive system being unable to flexibly adjust those expectations in the face of evidence of failure coming in from the offline world. This ongoing failure eventually teaches the system to expect failure - to predict its own inefficacy in the world - which is precisely the scenario described by AIF accounts of depression.

Thus, social media can put us in a bind: either we somehow bring the world into line with our new expectations, which might involve drastic action, or we risk experiencing symptoms of depression, engendered by an influx of inaccurate evidence which renders our generative model inflexible and inaccurate.

4 Designing Addictive Digital Spaces

Of course, there's a more obvious way to alleviate any rising prediction error resulting from too much time online: spend less time online. For some of us this is easier said than done though, as mounting evidence supports the suspicion that social media can be addictive. A comprehensive 2015 review defined social media addiction as a disproportionate concern with and drive to use social media that impairs other areas of life and found that roughly ten percent of users exhibit symptoms of addiction [3]. Interestingly, this is around the same percentage of people who have problems with alcohol – but while the addictive hooks of alcohol and other drugs are relatively well understood and uncontroversial, those of behavioural addictions such as engagement with social media are still subject to debate [28]. Some researchers argue that there is in fact no such thing as internet addiction at all [49]. Again, the active inference framework holds the key

to understanding why we should view engagement with digital hyperstimulators as potentially addictive.

The AIF offers a new understanding of addiction as a derailment of the alignment between predictive systems and their environment. Life contains various kinds of rewards: sex, food, status, etc., but for the brain all that matters is reducing prediction error, bringing us closer into expected states across various timescales. Dopamine encodes and reinforces behaviours that seek and pursue prediction error reduction. According to the AIF, dopamine plays a central role in encoding precision expectations on action policies [20][43]. Dopamine driven precision allocation is determined in part, as we have seen above, by changes in the rate at which error is reduced [10][29][23]. Addictive substances directly impact dopaminergic systems, signalling to the brain that a far better than expected slope of error reduction has taken place [44][35], which in turn upregulates precision on those drug seeking and taking policies. Through repeated use, the system comes to expect that vertiginous slope of error reduction, and is increasingly confident that it can be achieved only through drug seeking and taking behaviours. When the intense slope of error reduction associated with addictive substances cannot be met in ordinary life, the system registers failure, which is felt as disappointment, frustration, or pain (and a subsequent lowering of confidence on those policies). This progressive upregulation of precision on drug-seeking and taking behaviours, and down-regulation of non-addiction related action policies, leads to a narrowing of the agent's niche: friend groups are formed from fellow users, money is funneled toward buying drugs, and time is spent (when not using) planning how to acquire drugs. All the while the agent's other concerns related to family, career, friends, hobbies, and health are increasingly neglected [32]. A vicious feedback loop emerges - as mounting prediction errors spread across the broader concerns of an addicted individual's life (as these concerns are increasingly ignored and abandoned), the addict is more attracted to the one area where they can feel like they are succeeding, namely drug use. It is this feedback loop, that recruits the predictive mechanisms of brain and body to reorganize the habits of the addict around drug seeking and taking behaviour, that underlies the true pathology of addiction.

Just like alcohol and other drugs, digital environments threaten to disrupt this balance between naturally occurring rewards and reward seeking behaviour. In his important book 'Your Brain on Porn' [48], Gary Wilson argues that internet pornography presents itself as dangerously rewarding, pointing out that in one evening, internet porn facilitates levels of sexual novelty which would have been unavailable to our ancestors across an entire lifetime: multiple tabs or windows, hundreds of different models, escalating fetishes, all conspire to have our reward circuitry screaming "wow, we're doing far better than we ever thought possible!", when in reality we're just staring at a screen, alone. The novelty is particularly enticing, as our brains are always seeking new ways of reducing error, novel strategies for doing better than expected [23]. Our brains register this as a huge resolution of uncertainty, and our reward circuitry in the brain goes into overdrive, reinforcing these particular reward seeking behaviours.

What pornography is to sex, social media platforms are to our intrinsic appetite for socialising. Engaging in meaningful interpersonal bonding draws on all the reward circuitry mentioned above: it feels good to socialise, and dopamine entrenches learning for successful social behaviours [31]. One major similarity between social media and pornography is that both take a naturally occurring reward (sex and social behaviour, respectively), engineer a powerful vehicle of carefully curated fantasy, and present it as an attainable and desirable reality. These presentations of 'better than real life' scenarios (e.g. carefully staged and filtered images; maximally exciting sexual encounters in pornography) are highly alluring for predictive agents always on the lookout for ways to improve. On social media - just as with online porn - high levels of novelty and excess mean that the reward system is kicked into overdrive. It's no wonder that a 2019 report found that the average teenager in the US now spends more than seven hours a day looking at a screen [45]. Through social media, hyperstimulation can work to reorganise our predictive model and restructure our habits: we wake up and reach for our phone, never leave home without it, and constantly feel drawn toward our phone even when in the company of friends.

One avenue of objection here might be to point out that it seems debatable to what extent social media captures the reality of offline social interactions, devoid as it is of many features of face-to-face communication, and therefore it should be unclear just how rewarding online social interactions actually are. In response to this, we can first return to the comparison with online pornography, which clearly lacks the same substantive character as real sex and relationships. Nevertheless, engagement with online porn has been shown to powerfully engage the brain's reward seeking machinery [38][1]. This returns us to the point made earlier that, according to the AIF, all rewards are fundamentally processed as prediction error minimization – crack cocaine or explicit imagery, it's all the same: the system learns to expect states where error is reduced in line with or better than expected. While social media certainly lacks the face-to-face nuance of real-world interaction, it nevertheless works hard to turbo charge many of its most gratifying aspects such as judgement, monitoring, and validation and positive feedback.

In order to see how social media takes these rewarding aspects of social interaction and hyper charges them, first notice how all digital space has the inherent quality of dissolving the temporal and spatial restraints which govern offline interaction, thereby - in the case of social media - facilitating an excess of novelty and validation which simply isn't available in the real world. Users can instantaneously exchange direct messages with people who may well be complete strangers, and when users get bored of the content they're currently interacting with, a quick swipe generates new, exciting, unpredictable content. These structural features – which deliberately elicit anticipatory states and facilitate near endless potential for novelty – is something that deflationary accounts of social media addiction often fail to emphasize.

However, the potentially addictive nature of social media platforms doesn't only emerge from an excess of carefully edited content and potentially massive

social feedback. It also emerges from a deliberately designed and carefully implemented functional architecture which draws on our knowledge about the brain's reward circuitry and established approaches in the gambling industry. In gambling what's so arousing (and habit-forming) is the anticipation of reward, or the expectation of an uncertain reward [25]. Of course, offline social interactions are often unpredictable too, in that we don't know when someone might contact us or interact with us in rewarding ways, but social media sites are engineered to compound this anticipation through gamification, in which features such as progression, points scoring and risk taking are introduced into a non-game setting. Social media gamifies social interaction, primarily through various highly interactive systems of 'likes', 'shares', 'upvotes', comments and so on, which apply to user created content. This feedback is the direct measure of the 'success' of a particular post, and allows for comparisons in popularity between posts and posters.

When the potentially enormous levels of social feedback do come, it isn't immediately communicated to the user. Rather, we receive notifications in the form of a shining button or exciting sound which delays the discovery of the precise nature of the incoming content. The simple act of pushing a button to reveal information has been shown to trigger arousal and compulsive behaviour, and newly developed features on smartphones add further layers of anticipation [2]. The 'swipe to refresh' feature of the Facebook app's news feed, for example, where users physically swipe the screen to generate a new stream of information, is a startlingly similar action to the pulling of a casino slot machine arm. In each case, users don't know for sure what kind of content will spring up until they swipe. This feature, coupled with the fact that Facebook's feed is now effectively infinite has led to the app being described as "behavioural cocaine" [3].

One final layer of anticipation comes through the use of a smartphone itself, compounding the intermittency arousal of feedback and interaction: We've been so conditioned by anticipation of smartphone buzzing that "phantom vibration syndrome" – the erroneous sensation of our phone vibrating – now affects %65 - %89 of people who use smartphones [42][13]. Crucially, these carefully engineered spikes in user anticipation mirror the anticipatory states known to underlie problematic gambling; in people who exhibit addictive gambling behaviour, dopamine response has been shown to be most pronounced during phases of high anticipation [22]. Rather than the reward itself, it's these highly arousing states of expectation of reward which have been shown to elicit the strongest dopaminergic response [25][33], and the designers of these digital platforms know this.

5 Conclusion

The active inference framework has, in recent years, come to change how we understand a range of psychological phenomena, including addiction and depression. In this paper, we've used the theoretical tools of active inference to enter into an ongoing debate about the ways in which social media - and digital environments more broadly - have the potential to negatively impact our

mental wellbeing. While deflationary accounts downplay the effects of the inherent design of digital environments, this active inference account adds weight to arguments that there are inherent features of digital technology that can have profound consequences for our wellbeing. These arguments may have a wide-ranging impact, given that these inherent features are deliberately implemented. As design guru Nir Eyal states, "Companies increasingly find that their economic value is a function of the strength of the habits they create" [15]. As it turns out then, the designers of social media, aiming to maximize engagement through design, may have a de facto interest in increasing the corrosive effect their platforms have on the mental health of users. Seen in this context, this emerging scientific picture may lend significant weight to arguments that we should take digital hyperstimulants seriously as a threat to our wellbeing, and to voices calling for changes to the way digital technology like social media is designed, operated, and regulated.

References

1. de Alarcón, R., de la Iglesia, J.I., Casado, N.M., Montejo, A.L.: Online porn addiction: What we know and what we don't—a systematic review. Journal of clinical medicine **8**(1), 91 (2019)
2. Alter, A.: Irresistible: Why you are addicted to technology and how to set yourself free. Vintage (2017)
3. Andersson, H.: Social media apps are "deliberately" addictive to users. BBC News **3** (2018)
4. Arab, K., Barasain, O., Altaweel, A., Alkhayyal, J., Alshiha, L., Barasain, R., Alessa, R., Alshaalan, H.: Influence of social media on the decision to undergo a cosmetic procedure. Plastic and Reconstructive Surgery Global Open **7**(8) (2019)
5. Badcock, P.B., Davey, C.G., Whittle, S., Allen, N.B., Friston, K.J.: The depressed brain: an evolutionary systems theory. Trends in Cognitive Sciences **21**(3), 182–194 (2017)
6. Barrett, L.F., Quigley, K.S., Hamilton, P.: An active inference theory of allostasis and interoception in depression. Philosophical Transactions of the Royal Society B: Biological Sciences **371**(1708), 20160011 (2016)
7. Clark, A.: Surfing uncertainty: Prediction, action, and the embodied mind. Oxford University Press (2015)
8. Clark, A.: Predictions, precision, and agentive attention. Consciousness and cognition **56**, 115–119 (2017)
9. Cramer, S., Inkster, B.: Statusofmind–social media and young people's mental health and wellbeing.[online]. royal society for public health (2017)
10. Van de Cruys, S.: Affective value in the predictive mind. MIND Group; Frankfurt am Main (2017)
11. Curtis, S.: Social media users feel 'ugly, inadequate and jealous'. The Telegraph https://www.telegraph.co.uk/technology/social-media/10990297/Social-media-users-feel-ugly-inadequate-and-jealous.html
12. Deane, G., Miller, M., Wilkinson, S.: Losing ourselves: Active inference, depersonalization, and meditation. Frontiers in Psychology **11**, 2893 (2020)
13. Drouin, M., Kaiser, D.H., Miller, D.A.: Phantom vibrations among undergraduates: Prevalence and associated psychological characteristics. Computers in Human Behavior **28**(4), 1490–1496 (2012)

14. Eldar, E., Rutledge, R.B., Dolan, R.J., Niv, Y.: Mood as representation of momentum. Trends in cognitive sciences **20**(1), 15–24 (2016)
15. Eyal, N.: Hooked: How to build habit-forming products. Penguin (2014)
16. Fabry, R.E.: Into the dark room: a predictive processing account of major depressive disorder. Phenomenology and the Cognitive Sciences **19**(4), 685–704 (2020)
17. Flood, R.: Insta sham: I spent £30k on surgery to look like an instagram filter but instead get compared to the 'purge' mask. The Sun https://www.thesun.co.uk/fabulous/14374803/man-spend-30k-look-instgram-filter-purge-mask/
18. Friston, K.: The free-energy principle: a unified brain theory? Nature reviews neuroscience **11**(2), 127–138 (2010)
19. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: a process theory. Neural computation **29**(1), 1–49 (2017)
20. Friston, K.J., Shiner, T., FitzGerald, T., Galea, J.M., Adams, R., Brown, H., Dolan, R.J., Moran, R., Stephan, K.E., Bestmann, S.: Dopamine, affordance and active inference. PLoS computational biology **8**(1), e1002327 (2012)
21. Gritters, J.: How instagram takes a toll on influencers' brains. The Guardian (2019)
22. Hegarty, C., Eisenberg, D.P., Kohn, P., Dreher, J.C., Masdeu, J., Ianni, A.M., Turner, N., Gregory, M.D., Berman, K.F.: Ventral striatal dopamine synthesis correlates with neural activity during reward anticipation. In: NEUROPSYCHOPHARMACOLOGY. vol. 39, pp. S197–S198. NATURE PUBLISHING GROUP MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND (2014)
23. Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K.J., Ramstead, M.J.: Deeply felt affect: The emergence of valence in deep active inference. Neural computation **33**(2), 398–446 (2021)
24. Hohwy, J.: The predictive mind. Oxford University Press (2013)
25. van Holst, R.J., Veltman, D.J., Büchel, C., van den Brink, W., Goudriaan, A.E.: Distorted expectancy coding in problem gambling: is the addictive in the anticipation? Biological psychiatry **71**(8), 741–748 (2012)
26. Hunt, E.: Faking it: how selfie dysmorphia is driving people to seek surgery. The Guardian **23**(02) (2019)
27. Joffily, M., Coricelli, G.: Emotional valence and the free-energy principle. PLoS computational biology **9**(6), e1003094 (2013)
28. Kardefelt-Winther, D., Heeren, A., Schimmenti, A., van Rooij, A., Maurage, P., Carras, M., Edman, J., Blaszczynski, A., Khazaal, Y., Billieux, J.: How can we conceptualize behavioural addiction without pathologizing common behaviours? Addiction **112**(10), 1709–1715 (2017)
29. Kiverstein, J., Miller, M., Rietveld, E.: The feeling of grip: novelty, error dynamics, and the predictive brain. Synthese **196**(7), 2847–2869 (2019)
30. Kiverstein, J., Miller, M., Rietveld, E.: How mood tunes prediction: a neurophenomenological account of mood and its disturbance in major depression. Neuroscience of Consciousness **2020**(1), niaa003 (2020)
31. Kopec, A.M., Smith, C.J., Bilbo, S.D.: Neuro-immune mechanisms regulating social behavior: dopamine as mediator? Trends in neurosciences **42**(5), 337–348 (2019)
32. Lewis, M.: Brain change in addiction as learning, not disease. New England Journal of Medicine **379**(16), 1551–1560 (2018)
33. Linnet, J.: Neurobiological underpinnings of reward anticipation and outcome evaluation in gambling disorder. Frontiers in behavioral neuroscience **8**, 100 (2014)

34. Linson, A., Parr, T., Friston, K.J.: Active inference, stressors, and psychological trauma: A neuroethological model of (mal) adaptive explore-exploit dynamics in ecological context. Behavioural Brain Research **380**, 112421 (2020)
35. Miller, M., Kiverstein, J., Rietveld, E.: Embodying addiction: a predictive processing account. Brain and cognition **138**, 105495 (2020)
36. Moss, R.: Instagram's scarlett london on being in the centre of a social media storm. Huffington Post https://www.huffingtonpost.co.uk/entry/there-is-a-real-
37. Narangajavana, Y., Fiol, L.J.C., Tena, M.Á.M., Artola, R.M.R., García, J.S.: The influence of social media in creating expectations. an empirical study for a tourist destination. Annals of Tourism Research **65**, 60–70 (2017)
38. Negash, S., Sheppard, N.V.N., Lambert, N.M., Fincham, F.D.: Trading later rewards for current pleasure: Pornography consumption and delay discounting. The Journal of Sex Research **53**(6), 689–700 (2016)
39. Parr, T., Friston, K.J.: Uncertainty, epistemics and active inference. Journal of the Royal Society Interface **14**(136), 20170376 (2017)
40. Paulus, M.P., Feinstein, J.S., Khalsa, S.S.: An active inference approach to interoceptive psychopathology. Annual review of clinical psychology **15**, 97–122 (2019)
41. Ramstead, M.J., Wiese, W., Miller, M., Friston, K.J.: Deep neurophenomenology: An active inference account of some features of conscious experience and of their disturbance in major depressive disorder (2020)
42. Rothberg, M.B., Arora, A., Hermann, J., Kleppel, R., St Marie, P., Visintainer, P.: Phantom vibration syndrome among medical staff: a cross sectional survey. Bmj **341** (2010)
43. Schwartenbeck, P., FitzGerald, T.H., Mathys, C., Dolan, R., Friston, K.: The dopaminergic midbrain encodes the expected certainty about desired outcomes. Cerebral cortex **25**(10), 3434–3445 (2015)
44. Schwartenbeck, P., FitzGerald, T.H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., Friston, K.: Optimal inference with suboptimal models: addiction and active bayesian inference. Medical hypotheses **84**(2), 109–117 (2015)
45. Siegel, R.: Tweens, teens and screens: The average time kids spend watching online videos has doubled in 4 years. The Washington Post (2019)
46. Smith, R., Kuplicki, R., Feinstein, J., Forthman, K.L., Stewart, J.L., Paulus, M.P., Khalsa, S.S., Investigators, T.., et al.: An active inference model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. medRxiv (2020)
47. Truly: Surgery transformed my face into an instagram filter — hooked on the look (2019), https://www.youtube.com/watch?v=JXEqVL6-ENY
48. Wilson, G.: Your brain on porn: Internet pornography and the emerging science of addiction. Commonwealth Publishing Richmond, VA (2014)
49. Yellowlees, P.M., Marks, S.: Problematic internet use or internet addiction? Computers in human behavior **23**(3), 1447–1453 (2007)

Ideas worth Spreading: A Free Energy Proposal for Cumulative Cultural Dynamics

Abstract. While there is a fast growing body of theoretical work on characterizing cumulative culture, quantifiable models underlining its dynamics remain scarce. This paper provides an active-inference formalization and accompanying simulations of cumulative culture in two steps: Firstly, we cast cultural transmission as a bi-directional process of communication that induces a generalized synchrony (operationalized as a particular convergence) between the internal states of interlocutors. Secondly, we cast cumulative culture as the emergence of accumulated modifications to cultural beliefs from the local efforts of agents to converge on a shared narrative.

Keywords: Active Inference, Generalized Synchrony, Communication, Cumulative Culture, Cultural Dynamics.

1 Introduction

Research on cultural dynamics focuses on the examination of fluctuations in cultural beliefs and practices and their evolution from a systems perspective. These dynamics consist of three processes that are typically studied separately: the introduction of novel beliefs and practices to a culture (i.e., innovation), the transmission of established beliefs and practices within a population (i.e., innovation diffusion), and their change in prevalence (Kashima, Bain & Perfors, 2019).

While there is a fast growing body of theoretical and empirical literature on the processes of cultural evolution (Aunger, 2001; Buskell, Enquist &Jansson, 2019; Bettencourt, Cintrón-Arias, Kaiser, & Castillo-Chávez, 2006; Creanza, Kolodny, & Feldman, 2017; Dawkins, 1993; Dean, Vale, Laland, Flynn, & Kendal, 2014; Dunstone & Caldwell, 2018; Enquist, Ghirlanda & Eriksson, 2011; Gabora, 1995; Heylighen & Chielens, 2009; Kashima, Bain, & Perfors, 2019; Richerson, Boyd & Henrich, 2010; Stout & Hecht, 2017; Weisbuch, Pauker & Ambady, 2009), quantitative models that are able to integrate different approaches and insights from multiple disciplines into unified, quantifiable interpretations of theory and empirical data are in rapidly growing demand (Creanza, Kolodny & Feldman, 2017).

This is particularly true for the mechanisms of social transmission, which have been especially reviewed under theoretical models (Aunger, 2001; Bettencourt, Cintrón-Arias, Kaiser & Castillo-Chávez, 2006; Dawkins, 1993; Gabora, 1995; Heylighen & Chielens, 2009; Kashima, Bain, & Perfors, 2019; Weisbuch, Pauker & Ambady, 2009) while mathematical models for cultural transmission remain scarce in this field. The term "cultural transmission" typically denotes the transference and spread of any particular fashion, ideology, preference, language or behavior within a culture (Creanza, Kolodny & Feldman, 2017). A prominent stream of quantitative models for cultural transmission are inspired by epidemiology, and convert models used for predicting the spread of a virus to formalize the spread of an idea (Bettencourt, Cintrón-Arias, Kaiser & Castillo-Chávez, 2006).

While the comparison of an idea to a virus has its benefits from a structural perspective, it implies the controversial notion that an idea is simply copied during its transmission through cultural exchange between individuals. This notion is not only

intuitively insufficient for a realistic portrayal of communication dynamics, but also conflicts with established theoretical models of transmission on these same grounds.

Current literature in cultural psychology indicates that rather than being simply duplicated during transmission, cultural beliefs and practices are modified through the active interpretation of each individual (Kashima et al. 2019). Another example for the discrepancy between quantitative epidemiology models for transmission and theory is taken from the psychology of communication. Research in this field suggests that communication is conditioned upon a mutual shared reality (Echterhoff, Higgins & Levine, 2009), or "common ground" (Clark & Brennan, 1991) between interlocutors. According to these theories, not only does cultural information change during communication, but (contradictory to the one-sided transmission of cultural information from "transmitter" to "receiver" that is implicit in epidemiological models) both interlocutors are active participants in generating this change. "Grounding" theories suggest that communication involves more than simply formulating a message and sending it off, but requires the mutual belief that what is being said will be understood by all parties.

Crucially, the notion that cultural information resists alterations during its transmission conflicts with a fundamental and particularly distinguished theory of cultural transmission: cumulative culture (Dunstone & Caldwell, 2018; Stout & Hecht, 2017). This approach to cultural evolution reflects the idea that cultural traits are gradually modified through transmission such that adaptive modifications accumulate over historical time (Dean, Vale, Laland, Flynn & Kendal, 2014). This theory operates from a basic assumption that transmission of cultural information naturally involves its modification in a way that fundamentally conflicts with the depiction of transmission under a disease spread formalisation.

The cumulative conceptualisation of modifications to cultural information is prominent in the literature and may be the most representative of genuine complexities underlying cultural dynamics. However, this triumph entails perhaps an inevitable downfall in that such a complex depiction of culture has proven exceptionally challenging to model in quantitative accounts (Buskell, Enquist, & Jansson, 2019). This paper provides an active-inference based quantitative account of cumulative culture as an accumulation of changes to cultural information over multiple transmissions.

2 Method

An emerging conclusion from the literature is that the term "transmission" for describing the spread of cultural information seems impoverished, as it leaves out the retention of cultural information. As implied by active inference and theoretical models of communication, the acquisition of cultural beliefs is as fundamental to the understanding of cultural information spread as their transmission. For this reason, we will henceforth be referring to what is known in the literature as cultural transmission as communication, or more technically- the local dynamics of cumulative culture.

2.1 Simulating the Local Dynamics of Communication

In our model, cultural transmission is cast as the mutual attunement of actively inferring agents to each other's internal belief states. This builds on a recent formalisation of communication as active inference (Friston & Frith, 2015) which resolves the problem of hermeneutics, (i.e., provides a model for the way in which people are able to understand each other rather precisely despite lacking direct access to each other's internal representations of meaning) by appealing to the notion of generalised synchrony as signaling the emergence of a shared narrative to which both interlocutors refer to. In active inference, this shared narrative is attained through the minimisation of uncertainty, or (variational) free energy when both communicating parties employ sufficiently similar generative models. We build on this to suggest that having sufficiently similar generative models allows communicating agents to recombine distinct representations of a belief (expressed as generative models) into one synchronised, shared model of the world. When we simulate the belief-updating dynamics between interacting agents, the cultural reproduction of a particular idea takes the form of a specific convergence between their respective generative models.

Under this theory, the elementary unit of heritable information takes the form of an internal belief state, held by an agent with a certain probability. When we simulate the belief-updating dynamics between interacting agents, a reproduced cultural belief is carried by the minds (or generative models) of both interlocutors as a site of cultural selection, where it may be further reproduced through the same process. Our simulations of communication involve two active inference agents with distinct generative models and belief claims that engage in communication over a hundred time steps.

2.2 Simulating the Global Dynamics of Cumulative Culture

Cultural beliefs and practices spread within a society through communication, a process which we have referred to as the local dynamics of cumulative culture. This description is appropriate because the accumulated outcomes of each (local) dyadic interaction collectively determine the degree to which an idea is prevalent in a culture. Moving from local communication dynamics to a degree to which an idea is prevalent in a cumulative culture is what we will refer to as the global dynamics of cumulative culture.

In our simulations of a cumulative culture, 50 active inference agents simultaneously engage in local dyadic communication as shown in our first simulation, such that 25 couples are engaged in conversation at every given time step. At the first time step, all agents have relatively similar belief states- referred to as the status quo. When we introduce an agent holding a divergent belief state to that of the status quo in the population, it propagates through it via pseudo-random engagements of agents in dialogue. In a simulated world of actively inferring agents, their individual mental (generative) models are slightly modified with every interlocutor they encounter, as their distinct representations converge to a shared narrative (Constant, Ramstead, Veissière, & Friston, 2019). The attunement of interlocutor's to each other's generative models on the microscale thus translates over time and with multiple encounters into collective free energy minimisation on the macroscale.

3 A Generative Model of Communication

In our simulations, agents attempt to convince each other of a cultural belief by utilising generative models that operate with local information only. For the establishment of such generative models, we will formulate a partially observed Markov decision process (MDP), where beliefs take the form of discrete probability distributions (for more details on the technical basis for MDP'S under an active inference framework, see Hesp 2019).
Under the formalism of a partially observed Markov decision process, active inference entails a particular structure. Typically, variables such as agent's hidden states (x, s), observable outcomes (o) and action policies (u) are defined, alongside parameters (representing matrices of categorical probability distributions).

3.1 Perceptual Inference

The first level of this generative model aims to capture how agents process belief claims they are introduced to through conversation with other agents. The perception of others' beliefs (regarded in active inference as evidence) requires prior beliefs(represented as likelihood mapping A1 about how hidden states (s1) generate sensory outcomes (o). Specifically, our agents predict the likelihood of perceiving evidence toward a particular expressed belief, given that this belief is "the actual state of the world". Parameterising an agent's perception of an interlocutor's expression of belief in terms of precision values can be simply understood as variability in agents' general sensitivity to model evidence. High precisions here correspond to high responsiveness to evidence for a hidden state and low precisions to low responsiveness to evidence. Precisions for each agent were generated from a continuous gamma distribution which is skewed in favour of high sensitivity to evidence on a population level (See figure 1: Perception).

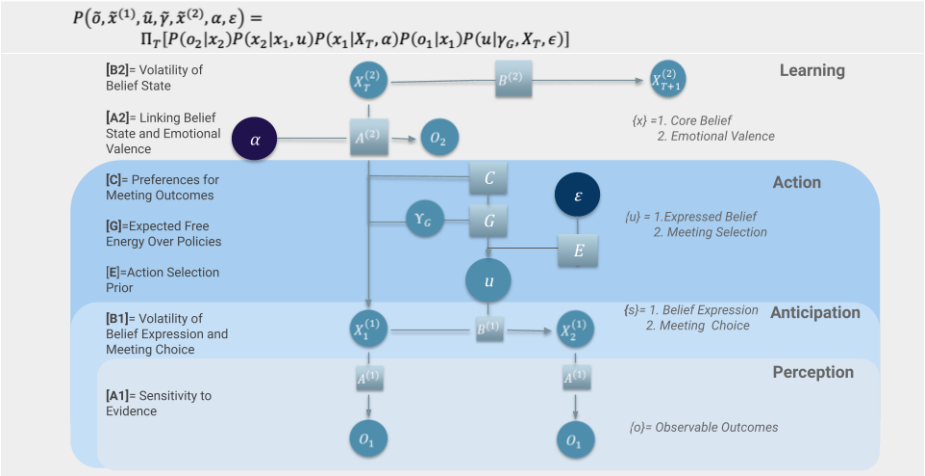


Fig. 1. A generative model of communication. Variables are visualised as circles, parameters as squares and concentration parameters as dark blue circles. Visualised on a horizontal line from left to right-states evolve in time. Visualised on a vertical line from bottom to top- parameters

build to a hierarchical structure that is in alignment with cognitive functions. Parameters are described to the left of the generative model and variables are described on the right.

Updating of core belief based on beliefs expressed by self and another agent after each meeting (detailed descriptions of the computations involved in perceptual inference can be found under appendix):

$$Q\left(x_{core}^{(2)}\right) = \sigma\left(\ln x_{core}^{(2)} + \gamma_{A,self}^{(2)} \ln o_{expr,self} + \gamma_{A,other}^{(2)} \ln o_{expr,other}\right) \quad (1)$$

3.2 Anticipation

At this level, our generative model specifies agents' beliefs about how hidden states (detailed in appendix A2) evolve over time. State transition probabilities [B1] define a particular value for the volatility of an agent's meeting selection
(s2) and belief expression (s1) [B1]. For each agent, this precision parameter is sampled from a gamma distribution, determining the a priori probability of changing state, relative to maintaining a current state. Note that belief states themselves are defined on the continuous range <0, 1> (i.e., as a probability distribution on a binary state), such that multiplication tends to result in a continuous decay of confidence over time in the absence of new evidence (where the rate of decay is inversely proportional to the precision on B) (See figure 1: Anticipation).

3.3 Action

After perceiving and anticipating hidden belief states in the world, our agents carry out deliberate actions biased towards the minimum of the expected free energy given each action (a lower level generative model for action is detailed in appendix A4 and A5). At each time point, a policy (U) is chosen out of a set of possible sequences for action. In our simulations, two types of actions are allowed: selecting an agent to meet at each given time point (u2) and selecting a specific belief to express in conversation (u1). The first allowable action holds 50 possible outcomes (one for each agent in the simulation) while the second is expressed on the range <0,1>, where the extremes correspond to complete confidence in denying or supporting the belief claim, respectively. Each policy under the G matrix specifies a particular combination of action outcomes weighted by its expected negative free energy value and a free energy minimising policy is chosen (See figure 1: Action).

Voluntary Meeting Selection. While the choice of interlocutor is predetermined in a dyad, our multi-agent simulations required some sophistication in formulating the underlying process behind agents' selection for a conversational partner (s3) at each of the hundred time points. Building on previous work on active inference navigation and planning (Kaplan & Friston, 2018), agents' meeting selection in our model is represented as a preferred location on a grid, where each cell on the grid represents a possible agent to meet (Appendix).

We demonstrate the feasibility of incorporating empirical cultural data within an active inference model by incorporating (1) confirmation bias through state-dependent

preferences [C], biasing meeting selection through the risk component of expected free energy (G) and (2) novelty seeking through the ambiguity component of expected free energy. The first form of bias reflects the widely observed phenomenon in psychology research that people's choices tend to be biased towards confirming their current beliefs (Nickerson, 1998). The second form of bias reflects the extent to which agents are driven by the minimisation of ambiguity about the beliefs of other agents, driving them towards seeking out agents they have not met yet.

3.4 Perceptual Learning

On this level agents anticipate how core belief states (specified in appendix A1) might change over time [B2] (figure 2.3). This is the highest level of cognitive control, where agents experience learning as a high cognitive function (higher level generative model is detailed in appendix A3). By talking with other simulated agents and observing their emotional and belief states, our agents learn associations between EV and beliefs via a high level likelihood mapping [A2], (updated via concentration parameter α). The Updating of core belief, based on beliefs expressed by other agents, is detailed in appendix A7. This learning is important because it provides our agents with certainty regarding the emotional value they can expect from holding the alternative belief to the status quo, which has low precision at the beginning of the simulation (before the population is introduced to an agent proclaiming this belief). The prior P(A) for this likelihood mapping is specified in terms of a Dirichlet distribution (Appendix).

4 Results

4.1 Local Dynamics of Coupled Communication

In nature, generalised synchrony emerges from a specific coupling between the internal states of dissipative chaotic systems (Pikovsky, Kurths, Rosenblum & Kurths, 2003). In active inference communication, agents are coupled in a bidirectional action-perception cycle in which they can be described as coupled dynamical systems (Friston & Frith 2015; Constant, Ramstead,Veissiere, Campbell, & Friston, 2018) Specifically, our model defines perceptual inference as the coupling parameter linking the internal states of interlocutors.
Also understood as sensitivity to model evidence (A1), perceptual inference is a direct and explicit form of coupling that occurs over the span of a single dialogue such that it modulates agents' convergence of internal belief states during conversation (Fig.2).
Our results indicate that without sufficiently high precisions on sensitivity to model evidence, agents' ability to listen and attune to the belief expression of their partner is limited to the extent that they are responsive to sensory evidence from their environment.

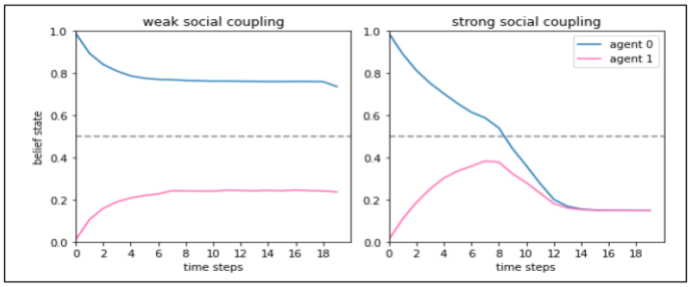


Fig.2: The [A1] parameter (sensitivity to model evidence) modulates the level of social coupling between agents in dialogue. **(Left)** When precision on sensitivity to model evidence is low (for both agents) their internal states are very weakly coupled, which results in each agent sticking to their own belief. **(Right)** When both agents have high sensitivity to model evidence, their beliefs converge into a shared representation of an idea that inhabits both of their generative models.

To get a sense of the implications of these simulations, it is important to make explicit the way in which they tie in to previous work on active inference communication. In 2015, Friston & Frith provided evidence for the notion that generalised synchrony becomes altogether unattainable when agents do not possess sufficiently similar generative models. Our model goes beyond this to provide evidence for the idea that only when generalised synchrony is attainable (i.e., when interlocutors possess sufficiently similar generative models), communication underlies a convergence between their belief states. Our simulations should therefore be understood as taking generalised synchrony for granted while providing evidence for the premise that the level to which agents' beliefs converge (i.e., the level of synchrony between their internal states) is modulated by their sensitivity to model evidence [A1].

4.2 Global Dynamics of Cumulative Culture

Our simulations of a cumulative culture should be understood as capturing the dynamics of a culture that is the sum (or-accumulation) of modifications to cultural beliefs and practices over time (Fig.3). While the local dynamics simulated in the previous section represent a single modification to cultural information (as a convergence between distinct belief states held by individual agents), these simulations accumulate these modifications and expose their emerging dynamics within the population. The fundamental achievement of these results is therefore their methodologically consistent and novel depiction of cumulative culture under a quantitative and measurable framework (namely, active inference).

We explain the communicative isolation observed in our simulations (Fig.3) as a self organised separation between groups of agents when they hold intractably divergent beliefs, such that communicative isolation best ensures local and collective free energy minimization. In other words, when an intractable divergent belief propagates within a homogenous population, communicative isolation between

incongruent groups emerges as a strategy to minimize expected free energy, while the same strategy homogenizes the belief states of agents within congruent groups.

The above simulations also show how changes to parameters that determine levels of confirmation bias [C] and novelty seeking [G] affect the segregation within the population into groups of agents holding either status quo congruent beliefs or the alternative belief. When novelty seeking is upregulated, the population evolves such that the majority of agents end up subscribing to the alternative belief. However, when confirmation bias is upregulated, the majority of agents end up subscribing to the status quo. What these results indicate is that novelty seeking on a local level stimulates the population as a whole toward the adoption of a belief that is divergent from the status quo. This happens when novelty seeking individuals, which are open and willing to meet with agents of unknown beliefs, are intrinsically encouraged by their own curiosity to engage with a divergent belief. Once such agents become gradually more favoring of this belief they start to popularize it to the rest of the population. If the population is however, populated by a vast majority of agents biased toward meeting individuals with confirming beliefs, they do not engage with an alternative belief and it does not get popularized.

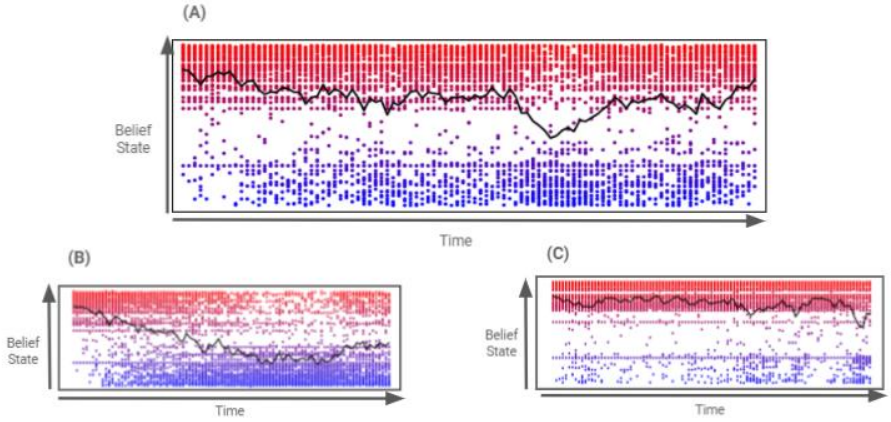


Fig 3: simulations of the spread of each agent's belief state (y) across time (x). 5o agents were used in this simulation and each of the 100 time steps represents the reproduced belief state outcomes of a particular combination of agents in dialogue. **(A)** Simulation of a Cumulative Culture. When a divergent belief state (blue) is introduced to the status quo population (red) at the first time step, it spreads through it via pseudo-random engagements of agents in dialogue that cumulatively change the belief structure within the population. Most notably, the introduction of a divergent belief seems to split the population into two subgroups: those holding a belief state that approximates the new divergent belief, and those holding an approximate status quo belief. This effect is modulated by agents' individual strategies for choosing which interlocutors to engage in conversation with (s3). **(B)** When novelty seeking is high in the population (above 10% of agents present high novelty seeking), the population is divided in favour of the divergent belief state, with more agents eventually holding this belief than the status quo. **(C)** When confirmation bias is high in the population (above 90% of agents present high confirmation bias) the population is divided in favour of the status quo belief, with more agents holding to this belief than the new and divergent belief.

5 Conclusion

In this paper, we employed an active inference model to tackle the complex task of formulating the dynamics underlying cumulative culture. Under this account, transmission is cast as a bidirectional process of communication that induces a generalised synchrony between the internal (belief) states of agents holding sufficiently similar generative models. Generalised synchrony is operationalised in our model as a particular convergence between the internal states of interlocutors, which is shown to be largely modulated by sensitivity to model evidence [A1].

When we simulate a population of agents that simultaneously engage in the converging dynamics of communication over time, cumulative culture emerges as the collective behavior brought about by these local modifications to cultural beliefs and practices. When a divergent belief is introduced to the status quo, it spreads within the population and brings about a collective behaviour that seems to be characterised by a divide between different belief groups. The level to which the status quo population defects to the divergent belief is mediated by local psychological strategies of confirmation bias and novelty seeking.

References

1. Aunger, R. (2001). Darwinizing culture: The status of memetics as a science.
2. Buskell, A., Enquist, M., & Jansson, F. . A systems approach to cultural evolution. Palgrave Communications, 5(1), 1-15 (2019).
3. Bettencourt, L. M., Cintrón-Arias, A., Kaiser, D. I., & Castillo-Chávez, C. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A: Statistical Mechanics and its Applications, 364, 513-536 (2006).
4. Clark, H. H., & Brennan, S. E. Grounding in communication (1991).
5. Constant, A., Ramstead, M. J., Veissière, S. P., & Friston, K. Regimes of expectations: An active inference model of social conformity and human decision making. Frontiers in psychology, 10, 679 (2019).
6. Constant, A., Ramstead, M. J., Veissiere, S. P., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. Journal of the Royal Society Interface, 15(141), 20170685.
7. Creanza, N., Kolodny, O., & Feldman, M. W. Cultural evolutionary theory: How culture evolves and why it matters. Proceedings of the National Academy of Sciences, 114(30), 7782-7789 (2017).
8. Dawkins, R. (1993). Viruses of the mind. Dennett and his critics: Demystifying mind, 13, e27.
9. Dean, L. G., Vale, G. L., Laland, K. N., Flynn, E., & Kendal, R. L. Human cumulative culture: a comparative perspective. Biological Reviews, 89(2), 284-301(2014).
10. Dunstone, J., & Caldwell, C. A. Cumulative culture and explicit metacognition: A review of theories, evidence and key predictions. Palgrave Communications, 4(1), 1-11 (2018).
11. Echterhoff, G., Higgins, E. T., & Levine, J. M. . Shared reality: Experiencing commonality with others' inner states about the world. Perspectives on Psychological Science, 4(5), 496-521 (2009).
12. Enquist, M., Ghirlanda, S., & Eriksson, K. (2011). Modelling the evolution and diversity of cumulative culture. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1563), 412-423.
13. Friston, K., & Frith, C. A duet for one. Consciousness and cognition, 36, 390-405 (2015).
14. Friston, K. J., & Frith, C. D. Active inference, communication and hermeneutics. cortex, 68, 129-143 (2015).
15. Gabora, L. (1995). Meme and variations: A computational model of cultural evolution. In 1993 Lectures in complex systems (pp. 471-485). Addison Wesley.
16. Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. . A multi-scale view of the emergent complexity of life: A free-energy proposal. In Evolution, Development and Complexity (pp. 195-227). Springer, Cham (2019) .
17. Kaplan, R., & Friston, K. J. Planning and navigation as active inference. Biological cybernetics, 112(4), 323-343 (2018).

18. Kashima, Y., Bain, P. G., & Perfors, A. The psychology of cultural dynamics: What is it, what do we know, and what is yet to be known?. Annual Review of Psychology, 70, 499-529 (2019)..
19. Pikovsky, A., Kurths, J., Rosenblum, M., & Kurths, J. Synchronization: a universal concept in nonlinear sciences (No. 12). Cambridge university press (2003).
20. Richerson, P. J., Boyd, R., & Henrich, J. (2010). Gene-culture coevolution in the age of genomics. Proceedings of the National Academy of Sciences, 107(Supplement 2), 8985-8992.
21. Stout, D., & Hecht, E. E. Evolutionary neuroscience of cumulative culture. Proceedings of the National Academy of Sciences, 114(30), 7861-7868 (2017).
22. Weisbuch, M., Pauker, K., & Ambady, N. (2009). The subtle transmission of race bias via televised nonverbal behavior. Science, 326(5960), 1711-1714.

Appendix A Generative Model Architecture, Factors and Parameters

A.1 Higher level hidden state factors:

$$x^{(2)} : \{x^{(2)}_{core},$$

[core beliefs of self and others about a particular claim, across days]

$$x^{(2)}_{mem},$$

[memory of having visited each agent]

$$x^{(2)}_{habit} \}$$

[habits of self, across days]

A.2 Lower level hidden state factors (specify events on a given 'day')

$$x^{(1)} : \{ \; x^{(1)}_{loc},$$

[self location, where each agent has a unique 'home' location]

$$x^{(1)}_{belief},$$

[beliefs of self and others about a particular claim]

$$x^{(1)}_{visit},$$

[beliefs about having visited each agent]

$$x^{(1)}_{sat} \; \}$$

[satisfaction of self and others]

A.3 Higher level generative model:

$$x^{(1)}_{belief} = A^{(2)}_{belief} x^{(2)}_{core}$$

[core beliefs specify prior expectations for beliefs on the lower level]

$$x^{(1)}_{sat} = A^{(2)}_{sat} x^{(2)}_{core}$$

[core beliefs specify satisfaction states for the lower level]

$$x^{(1)}_{visit} = A^{(2)}_{mem} x^{(2)}_{mem}$$

[memory specifies beliefs about having visited each agent on the lower level]

$$E_{expr} = A^{(2)}_{expr} x^{(2)}_{habit}$$

[habits of self specify prior tendency for belief expression]

$$x^{(2)}_{T+1} = B^{(2)} x^{(2)}_{T}$$

[higher-level states decay over time: gradual forgetting]

A.4 Lower level generative model for action:

Action model for meeting selection:

In our simulations, we have incorporated psychological biases in agents' preferences for meeting similar (i.e., belief compatible) or unknown agents. Note that while agents biased toward confirming beliefs would tend toward individuals with similar beliefs to their own, novelty seekers would not look for the opposite of this (i.e look for individuals of divergent beliefs to their own), but rather have a preference for individuals of yet unknown beliefs.

In active inference, action selection is guided by the expected free energy [G], which entails maximising the expected benefit or utility of an action (known as pragmatic value), while also maximising the potential information gain of future actions by reducing uncertainty about the causes of valuable outcomes (known as epistemic value). These constraints to action selection could be interpreted as formalising the exploration-exploitation trade-off in learning systems. Epistemic value (exploration) refers to the benefit related to searching over a sample space in order to get a better estimation of promising areas that will maximise pragmatic value (exploitation). Active-inference agents would therefore maximise epistemic value until information gain is low, after which the maximisation of pragmatic value and exploitation are assured (Friston, Rigoli, Ognibene, Mathys, Fitzgerald & Pezzulo, 2015).

In our model, agents' choice in meeting interlocutors with known and similar beliefs versus those with unknown beliefs can be cast in terms of a tradeoff between pragmatic and epistemic value. On the one hand, a confirmation bias emerges from the maximisation of expected utility, increasing synchronisation between interlocutors' internal models, thus allowing for the emergence of shared expectations (Hesp et al., 2019). On the other hand, novelty seeking emerges from the maximisation of information gain, allowing for the exploration of the sample space. Also understood as intrinsically motivated curious behaviour (Friston, Lin, Frith, Pezzulo, Hobson & Ondobaka, 2017), maximisation of epistemic value allows individuals to better predict the consequences of their actions (e.g.,, which agent to meet) through greater certainty about the hidden states of their environment (e.g., the beliefs of other agents).

From the point of view of agents in our simulations, increasing pragmatic value translates into selecting to meet interlocutors with similar beliefs, while increasing epistemic value translates into selecting agents whose beliefs are unknown or highly uncertain (This way, a meeting increases information gain). From this point of view, it is clear the two values constrain each other and maximizing both simultaneously is partially (but not entirely) paradoxical. While maximising pragmatic value requires agents to choose to meet with an interlocutor they know is similar to them, maximising epistemic value necessitates they meet with one they do not know at all.

$$P(u_{loc}) = \sigma\left(-\gamma_{G,loc}G_{loc} + \gamma_{E,loc}E_{loc}\right)$$

$$G_{loc} = o_{u,belief} \cdot (\ln o_{u,belief} - C_{belief}) + H \cdot x_{u,2,visit}$$

$$x_{u,2}^{(1)} = B_u^{(1)} x_1^{(1)}$$

$$o_{u,belief} = A_{belief}^{(1)} x_{u,2}^{(1)}$$

$$C_{belief} = \ln\left(A_C^{(2)} x_{core}^{(2)}\right)$$

$$if\ x_{visit,j} = 1:$$

[equals 1 if agent visited a particular agent *j*]

$$H_j = 0$$

[ambiguity is zero if agent visited this agent *j* already] $$else:$$

$$H_j = 0.1$$

[ambiguity is non-zero if agent has not visited agent *j* yet]

A6. Generative process:

Generative process for meeting selection:

$$u_{loc} \sim P(u_{loc})\ [actual\ meeting\ u_{loc}\ is\ sampled\ from\ meeting\ selection\ prior\ P(u_{loc})]$$

Generative process for belief expression and EV (satisfaction) of each agent:

At a high level of cognitive control, agents incorporate a series of processes underlying the selection of a particular belief for expression (u2). Other than the partial reliance on a low level habitual factor [E], this action involves multiple higher order considerations.

First, an agent considers their core belief state (x), and the way this state apriori maps on to one of two discrete emotional valence states (s2) via an initial likelihood mapping [A2] Emotional Valence (EV) is defined as the extent to which an emotion is positive or negative (Feldman Barrett & Russell, 1999), such that agents' core beliefs are apriori associated with either positive emotional valence or negative emotional valence (with some probability). As a minimal form of vicarious learning, the initial mapping is further updated based on associations agents observe between their interlocutors' expressed belief state and EV value The initial mapping therefore involves minimal precision for the expected EV for belief 2, since agents are first introduced to this belief (and associated EV) during the simulations. For this reason, the initial likelihood mapping between states is updated throughout our simulation via a crucial concentration parameter (α).

EV states are generated from core belief states, using a (learnable) likelihood mapping:

$$x_{sat}^{(1)} = A_{sat}^{(2)} x_{core}^{(2)}$$

Confidence of belief expression is generated using a Gamma distribution, where the rate parameter expris the Bayesian model average of (+,-)values associated with high and low satisfaction:

$$P\left(\gamma_{expr}\right) \approx \Gamma\left(1, \beta_{expr}\right)$$
$$\beta_{expr} = \beta^{(+,-)} \cdot x_{sat}^{(1)}, \qquad \beta^{(+,-)} = [0.25, 2.0]$$

The expression of beliefs is guided by current core beliefs (scaled with satisfaction-dependent expr) and by habitual belief expression Eexpr(scaled with a fixed parameter E,expr):

$$P\left(u_{expr}|\gamma_{expr}\right) = \sigma\left(-\gamma_{expr} \ln x_{core}^{(2)} + \gamma_{E,expr} E_{expr}\right)$$

The intrinsically stochastic and itinerant nature of the generative process of communication is modeled by using a two-dimensional Dirichlet distribution to generate observed expressions on
the range [0,1], where each agent's belief expression prior Puexpr|expr is used to specify their concentration parameters (multiplied by 12 to reduce variance):

$$o_{expr} = Dir(12u_{expr})$$

Generative process for emotional valence expressed by each agent:

$$o_{sat} = A_{sat}^{(1)} x_{sat}^{(1)}$$

[satisfaction observed by interaction partner corresponds to actual satisfaction]

The EV state predicted is then used to generate an action confidence value (γ) such that positive EV generates high confidence in a certain expression of the belief state (u1) and negative EV generates low confidence values. Higher confidence values produce higher precision on the expected free energy (G) for one's belief expressed in the current conversation.

A7. Perception:

Updating beliefs about the other agent's belief based on their expression:

$$Q\left(x_{belief}^{(1)}\right) = o_{expr}$$

Updating of core belief based on beliefs expressed by other agents:

$$Q\left(x_{core}^{(2)}\right) = \sigma\left(\ln ln\, x_{core}^{(2)} + \gamma_{A,self}^{(2)} \ln ln\, o_{expr,self} + \gamma_{A,other}^{(2)} \ln ln\, o_{expr,other}\right)$$

A8. Learning:

Habit learning for meeting selection:

$$P\left(E_{loc}\right) = Dir\left(e_{loc}\right)$$
$$Q\left(E_{loc}\right) = Dir\left(e_{loc} + 0.05u_{loc}\right)$$

Habit learning for belief expression:

$$P\left(E_{expr}\right) = Dir\left(e_{expr}\right)$$
$$Q\left(E_{expr}\right) = Dir\left(e_{expr} + 0.1o_{expr}\right)$$

Perceptual learning for the mapping between satisfaction and core beliefs, based on the expressions of other agents:

$$P\left(A_{sat}^{(2)}\right) = Dir\left(a_{sat}^{(2)}\right)$$

$$Q\left(A_{sat}^{(2)}\right) = Dir\left(a_{sat}^{(2)} + \gamma_{A}^{(2)} o_{expr} \ln ln\, x_{sat}^{(1)}\right)$$

A9. Initialisation of parameters for each agent:

$$\gamma^{(2)}_{A,belief} \sim \Gamma(5, 6)$$

[regulates the integration of beliefs of other agents in one's own core belief]

$$\gamma^{(2)}_{A,sat} \sim \Gamma(10, 1)$$

[regulates learning rate of mappings between satisfaction and core belief, based on observed correspondences in other agents]

$$\gamma_{G,loc} \sim \Gamma(1, 1)$$

[regulates reliance on action model in selecting agent to meet]

$$\gamma_{E,loc} \sim \Gamma(1, 1)$$

[regulates reliance on habitual prior in selecting agent to meet]

$$\gamma_{E,expr} \sim N\left(\frac{\gamma_{E,loc}}{10}, \frac{\gamma_{E,loc}}{200}\right)$$

[regulates reliance on habitual prior in expressing action, which correlates with $\gamma_{E,loc}$*]*

$$\gamma^{(2)}_{B,core} \sim \Gamma(4, .5)$$

[regulates stability of core beliefs across days]

$$\gamma^{(2)}_{B,habits} \sim \Gamma(.5, 1)$$

[regulates stability of expression habits across days]

$$B^{(2)}_0 = [[.75, .25], [.25, .75]]$$

[specifies baseline transition probabilities]

$$B^{(2)} = \sigma\left(\gamma^{(2)}_B \ln B^{(2)}_0\right)$$

[corrects $B^{(2)}_0$ *using the agent-specific* $\gamma^{(2)}_B$ *values]*

Agents with relatively weak confirmation bias:

$$A^{(2)}_C \sim Dir(6, 4)$$

[induces weak reliance on core beliefs for specifying lower-level preferences]

Agents with relatively strong confirmation bias:

$$A^{(2)}_{C,1} \sim Dir(999, 1)$$

[induces strong reliance on core beliefs for specifying lower-level preferences]

Dream to explore: 5-HT2a as adaptive temperature parameter for sophisticated affective inference

Adam Safron[1] and Zahra Sheikhbahaee[2]

[1] Center for Psychedelic and Consciousness Research, Department of Psychiatry & Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21224, USA

[2] David R. Cheriton School of Computer Science, University of Waterloo, ON, Canada

asafron@gmail.com

zsheikhb@uwaterloo.ca

Abstract. Relative to other neuromodulators, serotonin (5-HT) has received far less attention in machine learning and active inference. We will review prior work interpreting 5-HT1a signaling as an uncertainty parameter with opponency to dopamine. We will then discuss how 5-HT2a receptors may promote more exploratory policy selection by enhancing imaginative planning (as sophisticated affective inference). Finally, we will briefly comment on how qualitatively different effects may be observed across low and high levels of 5-HT2a signaling, where the latter may help agents to change self-adversarial policies and break free of maladaptive absorbing states in POMDPs.

Keywords: Serotonin, 5-HT1a, 5-HT2a, Sophisticated Active Inference, Affective Inference, Exploration, Exploitation, Imagination, Planning, Consciousness.

1 Introduction

Serotonin (5-HT) is a phylogenetically ancient monoamine neuromodulator found in all life forms, and which in mammals involves at least 14 distinct receptors that can be subdivided into 7 sub-classes [1]. This divergence of 5-HT systems arose through a process of evolutionary divergence via gene duplication and subsequent specialization of receptor subtypes and associated pathways [2]. This diversity may seem to suggest limited utility for attempting to recapitulate 5-HT-related functions in artificial systems. However, we propose that a substantial portion of 5-HT-related functionality may be obtained by focusing on the 1a and 2a receptors. We suggest this seemingly myopic focus on these two receptor classes may constitute a fruitful research direction on account of their highly conserved status in evolution, their relatively broad distribution in mammalian brains, as well as the common organismic significance of 5-HT signaling implicated by the mass-release of these diffusely acting neuromodulators from concentrated neurons in midbrain nuclei. While understanding the full diversity of 5-HT signaling will likely be illuminating with respect to abilities to differentially modify

various neural processes—with potentially common organismic significances—we believe a focus on 1a and 2a receptors will provide both maximal explanatory purchase and a foundation upon which subsequent modelling may proceed.

Cortex is highly populated by both 5-HT1a and 5-HT2a receptors [1]. 5-HT1a receptors have primarily inhibitory effects on neurons, and are functionally associated with somewhat subtle effects on mood, uncertainty, and the learning of complex behaviors, whereas 5-HT2a receptors have primarily excitatory effects and appear to have more pronounced effects on affect and cognitive processes [3]. 5-HT2a receptors also mediate the primary mechanism of action for hallucinogenic drugs such as lysergic acid diethylamide (LSD), psilocybin, and N,N-dimethyltryptamine (DMT) [4]. These compounds are widely known for inducing states of altered perception, thought, and feeling, with similarities to lucid dreaming; in this way, psychedelic states share features with both dreaming and waking consciousness [5].

The functionality of 5-HT1a and 5-HT2a receptors has respectively been associated with either passive or active coping strategies in the face of threat (or uncertainty) [6]. In this view, 5-HT1a signaling enables adaptive responses to mild-to-moderate stress through affective regulation and the inhibition of (disinhibitory) dopaminergic processes. 5-HT2a signaling, in contrast, is upregulated during more intense states of challenge—potentially including uncertainty with respect to highly valued goals—so allowing for both increased behavioral flexibility and neural plasticity [7].

The Free Energy Principle and Active Inference (FEP-AI) framework characterizes organisms as kinds of generative models [8], [9], with brains functioning as cybernetic control systems for embodied agents as they attempt to minimize uncertainty with respect to realizing their goals. These (both implicitly and explicitly) valued goals are understood as Bayesian prior preferences over likely outcomes that allow such systems to maintain their forms on evolutionary and developmental timescales. Towards this end, an FEP-AI agent maximizes model evidence for its existence by minimizing expected (variational) free energy (i.e., cumulative precision-weighted prediction errors) between its world model and sensory observations. This expected free energy minimization (and thereby self-model-evidence maximization) is realized either via perception (updating world models) or action (enactively updating world states). According to hierarchical predictive processing (HPP) models, cortical processes—and potentially biological systems more generally—continuously generate top-down predictions of bottom-up information at multiple levels of hierarchical abstraction. Notably, each level of these hierarchical generative models has varying levels of spatial and temporal granularity with respect to the latent system-world states it attempts to predict and alter through active inference, so allowing for multi-scale models with varying degrees of temporal depth and counterfactual richness [10]. In HPP, bottom-up observations are (efficiently) encoded as prediction errors, which ascend to higher cortical levels to update generative models when not predictively suppressed by top-down prior expectations. At the highest levels of abstraction in cortical hierarchies, maximal explanatory power may be found through models related to complex processes such as those underlying various forms of selfhood and self-consciousness [11]–[13].

Within FEP-AI, conscious planning is understood as “sophisticated active inference,” in which agents generate imagined sequences of counterfactual outcomes

through rolling out mental simulations of different patterns of action/policy selection [14], [15]. This imaginative planning takes the form of a deep tree search over counterfactual observations and actions, where different rollouts of simulated actions allow for exploring different branches of decision trees. This sophisticated active inference is governed by a singular objective function(al) of expected free energy, which achieves balance with respect to exploration-exploitation tradeoffs by selecting governing models that neither overfit nor underfit patterns of data in shaping perception and action. At perhaps the highest level of organization, patterns of action/policy selection via counterfactual processing (*e.g.*, simulated movements through space) are orchestrated by the hippocampal/entorhinal system [16], [17], where these spatiotemporal trajectories may be understood as constituting the stream of consciousness [13].

Below we describe parts of our ongoing explorations of ways in which the functional significances of 5-HT1a and 5-HT2a receptors may be understood through the lens of FEP-AI. If accurate, these models may provide a unified account of the roles of 5-HT in adaptive behavior with implications for machine-learning, neuropsychology, and evolutionary-developmental biology. We will characterize ways in which both 5-HT1a and 5-HT2a receptors influence the degrees to which agents initiate imaginative planning and offline learning via mental simulations, as opposed to more immediately releasing policies for overt goal-seeking behaviors. Finally, we will address some potential misconceptions about the functionality of different levels of 5-HT2a signaling, and further establish potentially fruitful connections to meta-reinforcement learning.

2 5-HT1a Receptors

5-HT1a receptors are found in different layers of the cortex, but they are most strongly expressed in layers V and VI [1]. 5-HT1a receptors suppress pyramidal cell activity by increasing rectifying K^+ currents, and have also been found to inhibit gamma oscillations in the hippocampus [18], potentially indicating reduced sensitivity to overall organismic prediction error [19]. This is opposite to the effects for DA (and in some respects 5-HT2a) signaling, which in FEP-AI is understood as enhancing the precision of bottom-up prediction errors [20], [21], so promoting the sensitivity of behavioral response to rewards. 5-HT1a signaling, in contrast, would instead promote deliberation and patience with respect to policy deployment [22]–[24].

Substantial experimental evidence has demonstrated opponency between dopaminergic (DA) and serotoninergic (5-HT) signaling [25], [26], which appear to be respectively associated with situations characterized by either more appetitive or aversive states. For biological organisms, appetitive motivational systems encourage approach while aversive systems promote avoidance and withdrawal. However, this is not to say that serotonin creates aversion, but rather that 5-HT signaling tends to be enhanced for situations in which organisms experience stress and uncertainty with respect to their ability to achieve their goals [27], including the fundamental goal of

survival. This is consistent with modulation of 5-HT1a neurons in the dorsal raphe nucleus [28], since release of action policies associated with either more passive or active coping ought to be modulated by the expected value of different patterns of enaction. However, firing rates for serotonergic also correlate with uncertainty more generally, indicating sensitivity to surprising events irrespective of the value of particular rewards. Opponency is observed in 5-HT1a and DA systems in terms of mutual inhibition of release and differential shaping of modulated systems [29]. However, their interactions can also produce synergy, both in terms of differentially parameterizing the nature of imaginative planning (e.g. with more or less confidence), and also in terms of providing a dynamic tension via their opponency, since pursuing complex goals require capacities for flexible adaption in response to environmental changes [30]–[33].

5-HT1a receptors are found as somatodendritic autoreceptors in raphe nuclei, as well as in postsynaptic sites in neocortex, hippocampus, and other "limbic" structures such as the amygdala and homeostatic regulatory nuclei of the hypothalamus [1]. Through inhibiting excitatory neurotransmission, 5-HT1a autoreceptors can both help with passively coping with stressful events by attenuating prediction error, and also help to promote more adaptive behavior by providing more time for planning in the face of uncertain circumstances. This inhibition of overt behavior affords both an opportunity for being informed by more complex world modeling, as well as an opportunity for adaptively calibrating world models by imbuing imaginative rollouts with greater uncertainty (i.e., implicitly functioning as a temperature parameter) [34]. Taken together, these functions allow for more flexible behavior via planning, as well as enhanced exploration and policy generalization via imagining the pursuit of goals under uncertain and potentially challenging conditions. Functionally speaking, this would be extremely sensible for a parameter that tends to be elevated in not just stressful circumstances, but also in situations involving the satiation of organismic drives such as eating, and possibly social contact [35]–[38]. That is, once goals are realized, a shift from exploration to exploitation would be both an adaptive foraging strategy and proximate mechanism for lifelong learning. Further, the more passive behavior encouraged by attenuation of action readiness would likely also be an adaptive response for an agent facing potential threats from circumstances that may overwhelm its present control abilities.

Thus, in FEP-AI terms, 5-HT1a signaling would be understood as promoting sophisticated inference via imaginative rollouts of predicted (or postdicted) patterns of enaction under conditions of reduced precision over counterfactually-deployed policies [14]. This would correspond to an agent experiencing relatively lower levels of confidence while entertaining counterfactual policies, but also with reduced "affective charge" [15], which one would normally expect to be stronger in a negative direction under conditions of reduced certainty with respect to realizing prior preferences. In these ways, 5-HT1a signaling would promote adaptive responses to challenging (and potentially novel) environments by increasing tolerance with respect to uncertainty through stress moderation, so allowing for more flexible and sophisticated forms of cognition and behavior [22]–[24].

The effects of different levels of 5-HT1a signaling may have profound functional consequences. For example, the ascending serotonergic pathway from the dorsal raphe nucleus and its effects on the amygdala and frontal cortex may promote adaptive reshaping conditioned fear responses [39], [40]. The basolateral nucleus of the amygdala contributes to behavioral changes in the face of emotional events and associated stimuli, including in response to stressors such as social defeat and other fearful circumstances [41], [42]. However, activation of 5-HT1a postsynaptic receptors in the dorsal hippocampus and amygdala facilitate extinction of fear-conditioned behaviors, consistent with the proposed roles of serotonin in facilitating coping in the face of threat [39]. With respect to potentially synergistic interactions between neuromodulatory systems, DA may be understood as providing a learning rate signal that influences the degree to which an agent updates its predictions in response to novel experiences [43]. This learning rate would influence the degree to which reward prediction errors shape policy selection, which if excessive could result in impulsivity by having presently estimated rewards promote more reactive forms of policy selection [44]. 5-HT1a, in contrast, would promote modeling with greater temporal depth and counterfactual richness, providing opportunities for meta-learning with sensitivity to (and ability to adaptively cope with) uncertain circumstances [24]. Notably, with respect to pathological states such as the rumination associated with depression and the impulsive aggression associated with antisocial behavior, low 5-HT1a in the medial prefrontal cortex appears to be associated with more perseverative tendencies and reduced abilities to adapt to novel environments [45], [46].

3 5-HT2a Receptors

5-HT2a receptors are most strongly expressed in high-level association cortices [1], including the “default mode network” (DMN). The DMN is comprised of a set of brain regions exhibiting high metabolic activity during resting states (including sleep), and which also become deactivated during goal-directed cognition, in conjunction with upregulation of “task positive” brain areas [47]. Notably, DA signaling tends to shift activity in the direction of increased dominance by frontoparietal control networks, and 5-HT signaling tends to shift activity towards exhibiting greater DMN power [48]. Key nodes of the DMN include medial prefrontal and parietal cortices [49], [50], as well as the temporoparietal junction [51], [52], which together constitute key areas for imaginative simulations involving both self and other [53], [54], and which may also be essential for establishing minimum embodied selfhood and coherent subjective experience [13], [55]. These nodes become functionally and structurally connected in a gradual manner over the course of development [56], and may have been uniquely expanded in the course of human evolution [57]. With respect to capacities for imaginative planning, it is particularly notable that the DMN plays a central role in counterfactual mental simulations [58], [59]. The DMN is often considered to be at the top of the cortical hierarchy in FEP-AI [11], although evidence suggests a more complicated picture in which salience networks may be understood as constituting the highest levels of control [60]. Notably, 5-HT2a receptors are also particularly

concentrated in the anterior insula [61], a key node in networks for salience determination and goal prioritization [62].

5-HT2a receptors are also responsible for the neuropsychological effects of psychedelics, which have been shown to elevate the entropy of endogenous brain activity and potentially enhance the richness of both the level and contents of consciousness [63]–[67]. With the widely known "RElaxed Beliefs Under pSychedelics" (REBUS) model [68], 5-HT2a agonists are suggested to attenuate the precision of high-level prior beliefs, so flattening the curvature of free energy landscapes and enhancing sensitivity to novel observations [69]. This altered regime promotes the breakdown of the brain's usual hierarchical structure [70], so allowing for an "anarchic" state in which novel forms of communication are allowed between brain areas, so allowing for enhanced cognitive exploration and opportunities for updating of deep beliefs. In this REBUS regime, ascending prediction errors from hierarchically lower levels may reshape upper level (potentially excessively precise) priors, so allowing the agent to break free from overly rigid patterns of thought and behavior.

In machine learning terms, such maladaptive cognitive and behavioral habits could be viewed as constituting self-adversarial policies, potentially formed through histories of excessive (or premature) exploitation in policy selection [71]. The more entropic dynamics afforded by high levels of 5-HT2a agonism, however, may allow agents to escape from these self-undermining attractors and reach more desirable regions of policy space. In terms of active inference, such relaxation of deep beliefs would be understood as reducing dominance from the parameters that serve as priors for agent-based generative models [72], which if excessively concentrated may preclude patterns of policy selection that could allow for opportunities for updating [73]. Concentration of probability mass in Dirichlet parameters via iterative policy selection and learning may provide a model of personality formation, and so their potential updating under conditions of strong 5-HT2a agonism (or functional homologues) could also provide a model of the kinds of personality change that have been associated with psychedelics [74]. This would also provide a model for the generation of novel, and potentially more (and possibly excessive) creative modes of cognition for both biological and artificial agents [75].

However, more physiologically typical low-to-moderate levels of 5-HT2a agonism have been suggested to involve a strengthening of beliefs under psychedelics (i.e., SEBUS effects) [76], both on account of increased activity from deep pyramidal neurons encoding prior expectations (or predictions), as well as reduced activity from superficial pyramidal neurons encoding prediction errors. Under this kind of SEBUS regime, individuals may engage in counterfactual processing under conditions of intense salience (e.g. sophisticated inference with high affective charge), potentially including greater confidence in both imagination and action. This would be highly consistent with accounts of 5-HT2a signaling as entailing strategies for "active coping" in the face of uncertainty/threat [6]. Whether patterns of either simulated or overtly enacted policy selection are more exploitative or exploratory would depend on a multitude of both pre-existing and context-specific priors over preferred patterns of enaction (e.g. typical levels of curious engagement) [77]. Further, both 5-HT1a and 5-HT2a signaling have been shown to inhibit sharp-wave ripples [78], which may

correspond to hippocampal/entorhinal remapping events [79]. Such inhibition of resetting of (generalized) mapping and accompanying repertoires of operative policies may promote opportunities for imaginative planning via more extended rollouts [80]. However, while 5-HT1a signaling will tend to be associated with more passive forms of cognition and behavior as described above, 5-HT2a signaling could promote more proactive modes of engagement with elevated affect from enhanced sensitivity to interoceptive signals [36]. In this way, 5-HT2a would provide a flexible parameter for imaginative planning when systems face varying degrees of stress (or uncertainty) with respect to achieving their goals.

SEBUS effects may also occur alongside REBUS effects at moderate-to-high levels of 5-HT2a agonism [68], [76], with potentially further indirect strengthening of intermediate level beliefs associated with the perceptual synthesis underlying conscious experience. This conjunction of high levels of perceptual vividness with exploration of novel forms of cognition could provide the greatest opportunities for updating, which may be a crucially important intervention for systems suffering from self-adversarial modes of policy selection. That is, while more physiological levels of 5-HT2a signaling may afford more flexible and adaptive refinement of normal policies, very high levels of agonism may constitute a qualitatively different regime that could allow both biological and artificial systems to "change their mind" in profound ways capable of altering their overall character [81]. Whether such changes are beneficial or harmful to system performance will depend on a multitude of factors, with the "set and setting" of such interventions being of crucial importance for shaping the direction of future system evolution. Going forward, we are currently planning simulation experiments in which we will demonstrate how these principles may apply to (artificial) world-modeling active inferential agents.

Acknowledgements

We gratefully acknowledge partial funding support from the Waterloo-Huawei Joint Innovation Lab within the project "the Active Inferential Meta-Learning Engine".

References

[1] N. M. Barnes *et al.*, "International Union of Basic and Clinical Pharmacology. CX. Classification of Receptors for 5-hydroxytryptamine; Pharmacology and Function," *Pharmacol. Rev.*, vol. 73, no. 1, pp. 310–520, Jan. 2021, doi: 10.1124/pr.118.015552.

[2] I. Moutkine, E. L. Collins, C. Béchade, and L. Maroteaux, "Evolutionary considerations on 5-HT2 receptors," *Pharmacol. Res.*, vol. 140, pp. 14–20, Feb. 2019, doi: 10.1016/j.phrs.2018.09.014.

[3] G. Zhang and R. W. Stackman, "The role of serotonin 5-HT2A receptors in memory and cognition," *Front. Pharmacol.*, vol. 6, p. 225, 2015, doi: 10.3389/fphar.2015.00225.

[4] M. W. Johnson, P. S. Hendricks, F. S. Barrett, and R. R. Griffiths, "Classic psychedelics: An integrative review of epidemiology, therapeutics, mystical experience, and brain network function," *Pharmacol. Ther.*, vol. 197, pp. 83–102, May 2019, doi: 10.1016/j.pharmthera.2018.11.010.

[5] R. Kraehenmann, “Dreams and Psychedelics: Neurophenomenological Comparison and Therapeutic Implications,” *Curr. Neuropharmacol.*, vol. 15, no. 7, pp. 1032–1042, 2017, doi: 10.2174/1573413713666170619092629.

[6] R. Carhart-Harris and D. Nutt, “Serotonin and brain function: a tale of two receptors,” *J. Psychopharmacol. Oxf. Engl.*, vol. 31, no. 9, pp. 1091–1120, Sep. 2017, doi: 10.1177/0269881117725915.

[7] L.-X. Shao *et al.*, “Psilocybin induces rapid and persistent growth of dendritic spines in frontal cortex in vivo,” *Neuron*, vol. 0, no. 0, Jul. 2021, doi: 10.1016/j.neuron.2021.06.008.

[8] K. J. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, “Active Inference: A Process Theory,” *Neural Comput.*, vol. 29, no. 1, pp. 1–49, Jan. 2017, doi: 10.1162/NECO_a_00912.

[9] K. J. Friston, “The free-energy principle: a unified brain theory?,” *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, Feb. 2010, doi: 10.1038/nrn2787.

[10] K. J. Friston, R. Rosch, T. Parr, C. Price, and H. Bowman, “Deep temporal models and active inference,” *Neurosci. Biobehav. Rev.*, vol. 77, pp. 388–402, 2017, doi: 10.1016/j.neubiorev.2017.04.009.

[11] R. L. Carhart-Harris and K. J. Friston, “The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas,” *Brain J. Neurol.*, vol. 133, no. Pt 4, pp. 1265–1283, Apr. 2010, doi: 10.1093/brain/awq010.

[12] A. Safron, “The Radically Embodied Conscious Cybernetic Bayesian Brain: From Free Energy to Free Will and Back Again,” *Entropy*, vol. 23, no. 6, Art. no. 6, Jun. 2021, doi: 10.3390/e23060783.

[13] A. Safron, “An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation,” *Front. Artif. Intell.*, vol. 3, 2020, doi: 10.3389/frai.2020.00030.

[14] K. Friston, L. Da Costa, D. Hafner, C. Hesp, and T. Parr, “Sophisticated Inference,” Jun. 2020, Accessed: Jun. 18, 2020. [Online]. Available: https://arxiv.org/abs/2006.04120v1

[15] C. Hesp, A. Tschantz, B. Millidge, M. Ramstead, K. Friston, and R. Smith, “Sophisticated Affective Inference: Simulating Anticipatory Affective Dynamics of Imagining Future Events,” in *Active Inference*, Cham, 2020, pp. 179–186. doi: 10.1007/978-3-030-64919-7_18.

[16] H. C. Barron, R. Auksztulewicz, and K. Friston, “Prediction and memory: a predictive coding account,” *Prog. Neurobiol.*, p. 101821, May 2020, doi: 10.1016/j.pneurobio.2020.101821.

[17] O. Çatal, T. Verbelen, T. Van de Maele, B. Dhoedt, and A. Safron, “Robot navigation as hierarchical active inference,” *Neural Netw.*, vol. 142, pp. 192–204, Oct. 2021, doi: 10.1016/j.neunet.2021.05.010.

[18] A. Johnston, C. J. McBain, and A. Fisahn, “5-Hydroxytryptamine1A receptor-activation hyperpolarizes pyramidal cells and suppresses hippocampal gamma oscillations via Kir3 channel activation,” *J. Physiol.*, vol. 592, no. 19, pp. 4187–4199, Oct. 2014, doi: 10.1113/jphysiol.2014.279083.

[19] F. Mannella, K. Gurney, and G. Baldassarre, “The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis,” *Front. Behav. Neurosci.*, vol. 7, p. 135, 2013, doi: 10.3389/fnbeh.2013.00135.

[20] T. H. B. FitzGerald, R. J. Dolan, and K. J. Friston, “Dopamine, reward learning, and active inference,” *Front. Comput. Neurosci.*, vol. 9, Nov. 2015, doi: 10.3389/fncom.2015.00136.

[21] K. J. Friston, P. Schwartenbeck, T. FitzGerald, M. Moutoussis, T. Behrens, and R. J. Dolan, “The anatomy of choice: dopamine and decision-making,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 369, no. 1655, Nov. 2014, doi: 10.1098/rstb.2013.0481.

[22] R. J. Moran *et al.*, “The Protective Action Encoding of Serotonin Transients in the Human Brain,” *Neuropsychopharmacology*, vol. 43, no. 6, Art. no. 6, May 2018, doi: 10.1038/npp.2017.304.

[23] C. D. Grossman, B. A. Bari, and J. Y. Cohen, “Serotonin neurons modulate learning rate through uncertainty,” *bioRxiv*, p. 2020.10.24.353508, Oct. 2020, doi: 10.1101/2020.10.24.353508.

[24] Y. Ohmura *et al.*, “Disruption of model-based decision making by silencing of serotonin neurons in the dorsal raphe nucleus,” *Curr. Biol.*, vol. 31, no. 11, pp. 2446-2454.e5, Jun. 2021, doi: 10.1016/j.cub.2021.03.048.

[25] Y.-L. Boureau and P. Dayan, “Opponency Revisited: Competition and Cooperation Between Dopamine and Serotonin,” *Neuropsychopharmacology*, vol. 36, no. 1, Art. no. 1, Jan. 2011, doi: 10.1038/npp.2010.151.

[26] N. D. Daw, S. Kakade, and P. Dayan, “Opponent interactions between serotonin and dopamine,” *Neural Netw. Off. J. Int. Neural Netw. Soc.*, vol. 15, no. 4–6, pp. 603–616, Jul. 2002, doi: 10.1016/s0893-6080(02)00052-7.

[27] K. Doya, K. W. Miyazaki, and K. Miyazaki, “Serotonergic modulation of cognitive computations,” *Curr. Opin. Behav. Sci.*, vol. 38, pp. 116–123, Apr. 2021, doi: 10.1016/j.cobeha.2021.02.003.

[28] E. S. Bromberg-Martin, O. Hikosaka, and K. Nakamura, “Coding of task reward value in the dorsal raphe nucleus,” *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 30, no. 18, pp. 6262–6272, May 2010, doi: 10.1523/JNEUROSCI.0015-10.2010.

[29] S. Yagishita, “Transient and sustained effects of dopamine and serotonin signaling in motivation-related behavior,” *Psychiatry Clin. Neurosci.*, vol. 74, no. 2, pp. 91–98, 2020, doi: 10.1111/pcn.12942.

[30] S. C. Hayes, *A Liberated Mind: How to Pivot Toward What Matters*. Penguin, 2019.

[31] S. Atasoy, G. Deco, and M. L. Kringelbach, “Playing at the Edge of Criticality: Expanded Whole-Brain Repertoire of Connectome-Harmonics,” in *The Functional Role of Critical Dynamics in Neural Systems*, N. Tomen, J. M. Herrmann, and U. Ernst, Eds. Cham: Springer International Publishing, 2019, pp. 27–45. doi: 10.1007/978-3-030-20965-0_2.

[32] A. K. Davis, F. S. Barrett, and R. R. Griffiths, “Psychological flexibility mediates the relations between acute psychedelic effects and subjective decreases in

depression and anxiety," *J. Context. Behav. Sci.*, vol. 15, pp. 39–45, Jan. 2020, doi: 10.1016/j.jcbs.2019.11.004.

[33] R. T. Gerraty, J. Y. Davidow, K. Foerde, A. Galvan, D. S. Bassett, and D. Shohamy, "Dynamic flexibility in striatal-cortical circuits supports reinforcement learning," *J. Neurosci.*, pp. 2084–17, Feb. 2018, doi: 10.1523/JNEUROSCI.2084-17.2018.

[34] D. Ha and J. Schmidhuber, "World Models," *ArXiv180310122 Cs Stat*, Mar. 2018, doi: 10.5281/zenodo.1207631.

[35] J.-P. Voigt and H. Fink, "Serotonin controlling feeding and satiety," *Behav. Brain Res.*, vol. 277, pp. 14–31, Jan. 2015, doi: 10.1016/j.bbr.2014.08.065.

[36] O. R. Hjorth *et al.*, "Expression and co-expression of serotonin and dopamine transporters in social anxiety disorder: a multitracer positron emission tomography study," *Mol. Psychiatry*, pp. 1–10, Dec. 2019, doi: 10.1038/s41380-019-0618-7.

[37] A. Fotopoulou and M. Tsakiris, "Mentalizing homeostasis: the social origins of interoceptive inference–replies to Commentaries," *Neuropsychoanalysis*, vol. 19, no. 1, pp. 71–76, 2017.

[38] A. Ciaunica, A. Constant, H. Preissl, and A. Fotopoulou, "The First Prior: from Co-Embodiment to Co-Homeostasis in Early Life." PsyArXiv, Jan. 05, 2021. doi: 10.31234/osf.io/twubr.

[39] I. V. Pavlova and M. P. Rysakova, "Effects of Administration of Serotonin 5-HT1A Receptor Ligands into the Amygdala on the Behavior of Rats with Different Manifestations of Conditioned Reflex Fear," *Neurosci. Behav. Physiol.*, vol. 48, no. 3, pp. 267–278, Mar. 2018, doi: 10.1007/s11055-018-0560-1.

[40] P. Dayan and Q. J. M. Huys, "Serotonin in Affective Control," *Annu. Rev. Neurosci.*, vol. 32, no. 1, pp. 95–126, 2009, doi: 10.1146/annurev.neuro.051508.135607.

[41] L. Colyn, E. Venzala, S. Marco, I. Perez-Otaño, and R. M. Tordera, "Chronic social defeat stress induces sustained synaptic structural changes in the prefrontal cortex and amygdala," *Behav. Brain Res.*, vol. 373, p. 112079, Nov. 2019, doi: 10.1016/j.bbr.2019.112079.

[42] P. B. Badcock, C. G. Davey, S. Whittle, N. B. Allen, and K. J. Friston, "The Depressed Brain: An Evolutionary Systems Theory," *Trends Cogn. Sci.*, vol. 21, no. 3, pp. 182–194, Mar. 2017, doi: 10.1016/j.tics.2017.01.005.

[43] W. Schultz, "Neuronal Reward and Decision Signals: From Theories to Data," *Physiol. Rev.*, vol. 95, no. 3, pp. 853–951, Jul. 2015, doi: 10.1152/physrev.00023.2014.

[44] J. W. Dalley and J. P. Roiser, "Dopamine, serotonin and impulsivity," *Neuroscience*, vol. 215, pp. 42–58, Jul. 2012, doi: 10.1016/j.neuroscience.2012.03.065.

[45] N. C. Di Pietro and J. K. Seamans, "Dopamine and serotonin interactions in the prefrontal cortex: insights on antipsychotic drugs and their mechanism of action," *Pharmacopsychiatry*, vol. 40 Suppl 1, pp. S27-33, Dec. 2007, doi: 10.1055/s-2007-992133.

[46] H. Lu and Q. Liu, “Serotonin in the Frontal Cortex: A Potential Therapeutic Target for Neurological Disorders,” *Biochem. Pharmacol. Open Access*, vol. 6, no. 1, p. e184, Feb. 2017, doi: 10.4172/2167-0501.1000e184.

[47] E. Dohmatob, G. Dumas, and D. Bzdok, “Dark control: The default mode network as a reinforcement learning agent,” *Hum. Brain Mapp.*, vol. 41, no. 12, pp. 3318–3341, 2020, doi: 10.1002/hbm.25019.

[48] B. Conio *et al.*, “Opposite effects of dopamine and serotonin on resting-state networks: review and implications for psychiatric disorders,” *Mol. Psychiatry*, vol. 25, no. 1, pp. 82–93, Jan. 2020, doi: 10.1038/s41380-019-0406-4.

[49] P. Fransson and G. Marrelec, “The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis,” *NeuroImage*, vol. 42, no. 3, pp. 1178–1184, Sep. 2008, doi: 10.1016/j.neuroimage.2008.05.059.

[50] A. V. Utevsky, D. V. Smith, and S. A. Huettel, “Precuneus Is a Functional Core of the Default-Mode Network,” *J. Neurosci.*, vol. 34, no. 3, pp. 932–940, Jan. 2014, doi: 10.1523/JNEUROSCI.4227-13.2014.

[51] B. Baird, A. Castelnovo, O. Gosseries, and G. Tononi, “Frequent lucid dreaming associated with increased functional connectivity between frontopolar cortex and temporoparietal association areas,” *Sci. Rep.*, vol. 8, no. 1, p. 17798, Dec. 2018, doi: 10.1038/s41598-018-36190-w.

[52] M. S. A. Graziano, “The temporoparietal junction and awareness,” *Neurosci. Conscious.*, vol. 2018, no. 1, Jan. 2018, doi: 10.1093/nc/niy005.

[53] D. Hassabis, R. N. Spreng, A. A. Rusu, C. A. Robbins, R. A. Mar, and D. L. Schacter, “Imagine All the People: How the Brain Creates and Uses Personality Models to Predict Behavior,” *Cereb. Cortex*, vol. 24, no. 8, pp. 1979–1987, Aug. 2014, doi: 10.1093/cercor/bht042.

[54] A. Guterstam, B. J. Bio, A. I. Wilterson, and M. Graziano, “Temporo-parietal cortex involved in modeling one’s own and others’ attention,” *eLife*, vol. 10, p. e63551, Feb. 2021, doi: 10.7554/eLife.63551.

[55] C. G. Davey and B. J. Harrison, “The brain’s center of gravity: how the default mode network helps us to understand the self,” *World Psychiatry*, vol. 17, no. 3, pp. 278–279, Oct. 2018, doi: 10.1002/wps.20553.

[56] F. Fan *et al.*, “Development of the default-mode network during childhood and adolescence: A longitudinal resting-state fMRI study,” *NeuroImage*, vol. 226, p. 117581, Feb. 2021, doi: 10.1016/j.neuroimage.2020.117581.

[57] R. L. Buckner and L. M. DiNicola, “The brain’s default network: updated anatomy, physiology and evolving insights,” *Nat. Rev. Neurosci.*, vol. 20, no. 10, pp. 593–608, Oct. 2019, doi: 10.1038/s41583-019-0212-7.

[58] D. Hassabis and E. A. Maguire, “The construction system of the brain,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 364, no. 1521, pp. 1263–1271, May 2009, doi: 10.1098/rstb.2008.0296.

[59] L. Faul, P. L. St. Jacques, J. T. DeRosa, N. Parikh, and F. De Brigard, “Differential contribution of anterior and posterior midline regions during mental simulation of counterfactual and perspective shifts in autobiographical

memories," *NeuroImage*, vol. 215, p. 116843, Jul. 2020, doi: 10.1016/j.neuroimage.2020.116843.

[60] Y. Zhou, K. J. Friston, P. Zeidman, J. Chen, S. Li, and A. Razi, "The Hierarchical Organization of the Default, Dorsal Attention and Salience Networks in Adolescents and Young Adults," *Cereb. Cortex N. Y. NY*, vol. 28, no. 2, pp. 726–737, Feb. 2018, doi: 10.1093/cercor/bhx307.

[61] A. M. Santangelo *et al.*, "Insula serotonin 2A receptor binding and gene expression contribute to serotonin transporter polymorphism anxious phenotype in primates," *Proc. Natl. Acad. Sci.*, vol. 116, no. 29, pp. 14761–14768, Jul. 2019, doi: 10.1073/pnas.1902087116.

[62] A. R. Rueter, S. V. Abram, A. W. MacDonald, A. Rustichini, and C. G. DeYoung, "The goal priority network as a neural substrate of Conscientiousness," *Hum. Brain Mapp.*, vol. 39, no. 9, pp. 3574–3585, Sep. 2018, doi: 10.1002/hbm.24195.

[63] L. Barnett, S. D. Muthukumaraswamy, R. L. Carhart-Harris, and A. K. Seth, "Decreased directed functional connectivity in the psychedelic state," *NeuroImage*, vol. 209, p. 116462, Apr. 2020, doi: 10.1016/j.neuroimage.2019.116462.

[64] M. M. Schartner, R. L. Carhart-Harris, A. B. Barrett, A. K. Seth, and S. D. Muthukumaraswamy, "Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin," *Sci. Rep.*, vol. 7, p. 46421, Apr. 2017, doi: 10.1038/srep46421.

[65] J. Aru, M. Suzuki, R. Rutiku, M. E. Larkum, and T. Bachmann, "Coupling the State and Contents of Consciousness," *Front. Syst. Neurosci.*, vol. 13, Aug. 2019, doi: 10.3389/fnsys.2019.00043.

[66] R. L. Carhart-Harris, "The entropic brain - revisited," *Neuropharmacology*, vol. 142, pp. 167–178, Nov. 2018, doi: 10.1016/j.neuropharm.2018.03.010.

[67] R. L. Carhart-Harris *et al.*, "The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs," *Front. Hum. Neurosci.*, vol. 8, p. 20, 2014.

[68] R. L. Carhart-Harris and K. J. Friston, "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics," *Pharmacol. Rev.*, vol. 71, no. 3, pp. 316–344, Jul. 2019, doi: 10.1124/pr.118.017160.

[69] A. I. Luppi *et al.*, "Connectome Harmonic Decomposition of Human Brain Dynamics Reveals a Landscape of Consciousness," *bioRxiv*, p. 2020.08.10.244459, Aug. 2020, doi: 10.1101/2020.08.10.244459.

[70] A. I. Luppi, R. L. Carhart-Harris, L. Roseman, I. Pappas, D. K. Menon, and E. A. Stamatakis, "LSD alters dynamic integration and segregation in the human brain," *NeuroImage*, vol. 227, p. 117653, Feb. 2021, doi: 10.1016/j.neuroimage.2020.117653.

[71] K. O. Stanley and J. Lehman, *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, 2015.

[72] A. Safron and C. G. DeYoung, "Chapter 18 - Integrating Cybernetic Big Five Theory with the free energy principle: A new strategy for modeling personalities as complex systems," in *Measuring and Modeling Persons and Situations*, D.

Wood, S. J. Read, P. D. Harms, and A. Slaughter, Eds. Academic Press, 2021, pp. 617–649. doi: 10.1016/B978-0-12-819200-9.00010-7.

[73] A. Constant, C. Hesp, C. G. Davey, K. J. Friston, and P. B. Badcock, "Why Depressed Mood is Adaptive: A Numerical Proof of Principle for an Evolutionary Systems Theory of Depression," *Comput. Psychiatry*, vol. 5, no. 1, Art. no. 1, Jun. 2021, doi: 10.5334/cpsy.70.

[74] D. Erritzoe, J. Smith, P. M. Fisher, R. Carhart-Harris, V. G. Frokjaer, and G. M. Knudsen, "Recreational use of psychedelics is associated with elevated personality trait openness: Exploration of associations with brain serotonin markers," *J. Psychopharmacol. Oxf. Engl.*, p. 269881119827891, Feb. 2019, doi: 10.1177/0269881119827891.

[75] M. Girn, C. Mills, L. Roseman, R. L. Carhart-Harris, and K. Christoff, "Updating the dynamic framework of thought: Creativity and psychedelics," *NeuroImage*, vol. 213, p. 116726, Jun. 2020, doi: 10.1016/j.neuroimage.2020.116726.

[76] A. Safron, "Strengthened beliefs under psychedelics (SEBUS)? A Commentary on 'REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics.'" PsyArXiv, Nov. 30, 2020. doi: 10.31234/osf.io/zqh4b.

[77] P. Schwartenbeck, J. Passecker, T. U. Hauser, T. H. FitzGerald, M. Kronbichler, and K. J. Friston, "Computational mechanisms of curiosity and goal-directed exploration," *eLife*, vol. 8, 10 2019, doi: 10.7554/eLife.41703.

[78] R. ul Haq *et al.*, "Serotonin dependent masking of hippocampal sharp wave ripples," *Neuropharmacology*, vol. 101, pp. 188–203, Feb. 2016, doi: 10.1016/j.neuropharm.2015.09.026.

[79] P. Latuske, O. Kornienko, L. Kohler, and K. Allen, "Hippocampal Remapping and Its Entorhinal Origin," *Front. Behav. Neurosci.*, vol. 11, 2018, doi: 10.3389/fnbeh.2017.00253.

[80] C. O'Callaghan, I. C. Walpola, and J. M. Shine, "Neuromodulation of the mind-wandering brain state: the interaction between neuromodulatory tone, sharp wave-ripples and spontaneous thought," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 376, no. 1817, p. 20190699, Feb. 2021, doi: 10.1098/rstb.2019.0699.

[81] M. Pollan, *How to Change Your Mind: The New Science of Psychedelics*. Penguin Books Limited, 2018.

Inferring in Circles: Active Inference in Continuous State Space using Hierarchical Gaussian Filtering of Sufficient Statistics

Peter Thestrup Waade[1,2][0000−0002−6061−0084], Nace Mikus[1,3][0000−0002−3445−9464], and Christoph Mathys[1,4,5][0000−0003−4079−5453]

[1] Interacting Minds Centre (IMC), Aarhus University, Aarhus, Denmark
[2] Embodied Cognition Group (ECG), Aarhus University, Aarhus, Denmark
[3] Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Austria
[4] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[5] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland

Abstract. We create a continuous state space active inference agent based on the hierarchical Gaussian filter. It uses the HGF to track the sufficient statistics of noisy observations of a moving target that is performing a Gaussian random walk with drift and varying volatility. On the basis of this filtering, the agent predicts the target's position, and minimizes surprisal by staying close to it. Our simulated agent represents the first full implementation of this approach. It demonstrates the feasibility of supplementing active inference with HGF-filtering of the sufficient statistics of observations, which is particularly useful in noisy and volatile continuous state space environments.

Keywords: active inference, continuous state space, sufficient statistics filtering, precision-weighted prediction errors, hierarchical Gaussian filter

1 Introduction

Active inference [7] is a formal framework for programming and modelling agents that navigate their environment such that they sample evidence for being within a desired set of states. This is done by minimizing the surprisal of sensory observations relative to a generative model of the environment, in which preferences for states are encoded as prior expectations. Actions are then chosen that are expected to lead to less surprisal in the future. Evaluating surprisal exactly is usually computationally intractable. In practical implementations of active inference, a variational free energy approximation is therefore often used.

Active inference furnishes a modeling framework which unites action and perception under a shared optimization imperative. Models inherently include a balance between epistemic and pragmatic behavior [4], can be related to neurobiological theories such as predictive processing [6], and can be motivated from first principles in physics and information theory [8,2].

Recently, most active inference agents have been implemented as partially observable Markov Decision Processes (POMDP's) [18]. Here agents are limited to making discrete actions and observations in a discrete state space. By contrast, we here aim to (re-)extend active inference models to the continuous domain. We demonstrate a principled approach where an active inference agent filters the sufficient statistics of its observations with a hierarchical Gaussian filter [12,13], allowing it to perform goal-directed actions in a noisy and volatile continuous state space-environment.

2 Filtering Sufficient Statistics with Hierarchical Gaussian Filters

For agents inferring continuous states obscured by observational, informational, and environmental uncertainty, a fundamental challenge is to filter these various sources of noise from their observations. One principled way of solving this problem, which we use here and which is consistent with active inference in general, is to invert a generative model of what causes sensory observations. The hierarchical Gaussian filter's update equations implement such an inversion, where the generative model consists of a hierarchical cascade of random walks [13]. The update equations in the HGF are a more efficient alternative to variational Laplace, as detailed in [12]. Given a time series of observations, this allows for teasing apart observation noise, (potentially changing) volatility and (possibly state-dependent) regularities like drifts and biases. HGFs also provide precision-weighted predictions for future states, and can be used to infer a full predictive posterior probability distribution over observations in the future. This can be done by constructing the predictive distribution such that it reflects the uncertainty implied in the HGF's updates when filtering the sufficient statistics of the observations [14].

The decisive point here is that in a Gaussian model for a continuous univariate hidden state (i.e., Gaussian prior and Gaussian likelihood), the prior and posterior predictive distributions are *Gaussian-predictive* distributions, corresponding to a reparameterization of the generalized Student's-t distribution. This means that in addition to location and scale parameters, the predictive distribution also has a degree-of-freedom parameter which determines the fatness of its tails. An appropriate filter, such as an HGF, allows for inferring all three of these parameters. We put this to use here in order for our active inference agent to make the most appropriate predictions possible, i.e. predictions which minimize surprisal by optimizing all three of their aspects: point prediction (mean), uncertainty (variance), and fatness of tails (degrees of freedom). In the next section, we demonstrate how this can be accomplished in a simple active inference context.

3 Active inference in Continuous State Space

We here provide a framework for a proof-of-principle simulation with a simple HGF-based active inference agent. The agent's objective is to stay close to a target which moves in continuous space with varying volatility. We will here first describe the *generative process* that forms the environment, and how it interfaces with the agent's *observations* o and *actions* a. Then we will describe how the agent makes actions as active inference based on inferences from the HGF.

The generative process consists of a total of three hidden states. The first is the target's position x_1 which moves in a Gaussian random walk with a constant drift ρ:

$$x_{1,t} \sim \mathcal{N}(x_{1,t-1} + \rho, x_{2,t}) \tag{1}$$

The second is the volatility of the random walk x_2, which changes in a pre-specified pattern between low and high levels of volatility. The last is then the agent's own position x_{agent}. In this example simulation, the position is fully determined by the agent's action a_{agent}, implemented here as being sampled from a delta distribution:

$$x_{\text{agent},t} \sim \delta(a_{\text{agent},t}) \tag{2}$$

The target's position x_1 is observed noisily, with observations o_1 normally distributed around the true position with standard deviation σ:

$$o_{1,t} \sim \mathcal{N}(x_{1,t}, \sigma) \tag{3}$$

The agent also observes its own position perfectly:

$$o_{\text{agent},t} \sim \delta(x_{\text{agent},t}) \tag{4}$$

To make inferences and predictions about the position and volatility of the target, the agent uses a standard HGF with a single volatility parent and a drift on the position. On each timestep t, this gives the agent a Gaussian belief about the target's position on the next timestep with mean $\hat{\mu}_{x_1,t}$ and precision $\hat{\pi}_{x_1,t}$. This lets it generate a full predictive posterior probability distribution for the observation on the next timestep. This distribution is a t-distribution with $\nu_t + 1$ degrees of freedom, with location $\hat{\mu}_{x_1,t}$ and precision $\hat{\pi}_{x_1,t}$:

$$p_{PP}(o_{1,t+1}|o_{1,1:t}) = t\left(o_{1,t}|\hat{\mu}_{x_1,t}, \hat{\pi}_{x_1,t}, \nu_t + 1\right) \tag{5}$$

where

$$\nu_t = \frac{\hat{\pi}_{x_1,t}}{\pi_\epsilon}, \tag{6}$$

and π_ϵ is the agent's prior belief about the input precision. In addition, the agent is equipped with a static prior distribution encoding its expectations (i.e.

preferences) for observations. The *goal prior*, as it will be referred to, is here a probability distribution over differences between the observed position of the target o_1 and the observation of the agent's own position o_{agent}. Specifically, it is a Gaussian distribution, with the mean μ_{GP} (usually at 0) encoding the preferred position relative to the target and the precision π_{GP} specifying the strength of this preference:

$$p_{GP}(o_1 - o_{\text{agent}}) = \mathcal{N}(o_1 - o_{\text{agent}}; \mu_{GP}, \pi_{GP}) \tag{7}$$

On each trial, the agent's surprisal is calculated as the negative log probability of its sensory input relative to the goal prior:

$$\Im(o_1 - o_{\text{agent}}) = -\ln\, p_{GP}(o_1 - o_{\text{agent}}) \tag{8}$$

In order to choose its action, the agent creates an expected surprisal distribution over possible control states a. First the expectation of the predictive posterior over observations of the target is assumed as the observation of the target. This gives a time-varying distribution over the agent's preferences for observations of its own position, given that the target is observed at its expectation. In the agent's model of the environment, equations 2 and 4 are recapitulated, so we can substitute the expected observation o_{agent} with the agent's control states a_{agent}:

$$p_{GP,t}(a_{\text{agent}}) = p_{GP}(o_{\text{agent}}|o_1 = E(p_{PP}(o_1, t))) \tag{9}$$

The right side of this equation is the goal prior over observations of the agent, given that the target is observed at its expectation $E(p_{PP}(o_1, t))$. The left side of the equation $p_{GP,t}(a_{\text{agent}})$ then becomes what might be called a 'goal posterior', a distribution over preferences for actions. In order to incorporate the full uncertainty of the agent's predictions, however, this preference distribution is convolved with the full predictive posterior. Taking the negative log of the resulting probability distribution then yields the expected surprisal associated with each possible move, after including the full uncertainty:

$$\Im_{\text{expected},t}(a_{\text{agent}}) = -\ln\, p_{PP}(o_1) * p_{GP,t}(a_{\text{agent}}) \tag{10}$$

The agent then selects deterministically the action with the lowest associated expected surprisal.

$$a_{\text{agent},t} = \underset{a}{\text{argmin}} \quad \Im_{\text{expected},t}(a_{\text{agent}})) \tag{11}$$

4 Example Simulation

We here show results from an example simulation with the environment and the HGF-based active inference implementation described in the previous section. Figure 1 shows a schematic of the inference, prediction and decision process of the active inference agent. A GIF demonstrating the agent moving to follow the

noisy and volatile observations of the target can be found on this link: `https://osf.io/x5v39/`

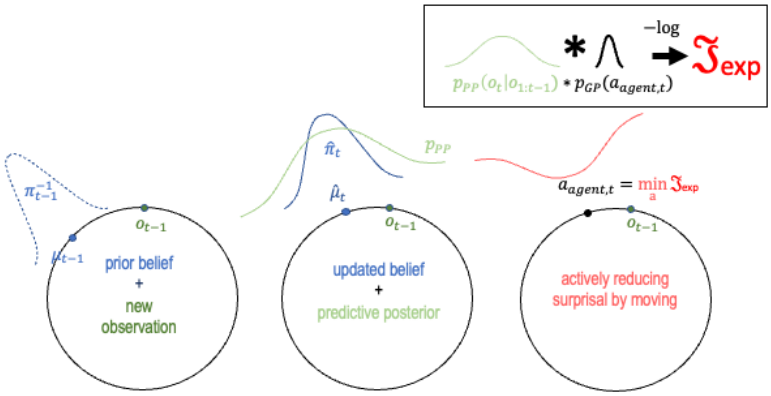


Fig. 1. Graphical sketch of the agent's action process, visualized on the circle. The agent starts with a Gaussian prior belief about the target's position with mean μ_{t-1} and precision π_{t-1}, and makes a new observation o_t. From that a new belief is computed with the HGF (also taking into consideration the drift ρ), and a predictive posterior t-distribution p_{PP} can be calculated. Finally, the predictive posterior is convolved with the 'goal posterior' $p_{GP}(a_{agent,t})$ i.e. the goal prior over agent positions given that the target is observed at its expectation (see equation 9). The negative log of the resulting probability distribution is the expected surprisal associated with the agent moving to different positions, of which the lowest is selected. If the static goal prior $p_{GP}(o_1 - o_{agent})$ is symmetric and centered around 0, μ_t, the mean of p_{PP} and the agent's action $a_{agent,t}$ coincide.

Figure 2 shows the inference on the target position, the prediction of future observations and the subsequent movement of the agent in an example simulation. Here the agent's goal prior is centered around 0, meaning that it consistently moves to the mean of its predictive posterior. Figure 3 shows the volatility of the environment and the agents inferred volatility in the same simulation. Note that this is a stochastic process, so even though the generating volatility is high it is not guaranteed that the target will move more. This means that the optimal inference on the volatility is not always the same as the volatility that generated the data (as in this case), although it should be when averaged over many simulations (this is also potentially true for any other hidden aspect of the environment). Figure 4 shows the agent's surprisal at its observation, relative to the goal prior. As expected, surprisal and predictive uncertainty is generally

higher in those periods where the actual volatility is higher. In general, the agent performs its task well, demonstrating the feasibility of using the HGF for performing active inference. Code to replicate and modify this simulation can be found on `https://github.com/ilabcode/hgf-active-inference`

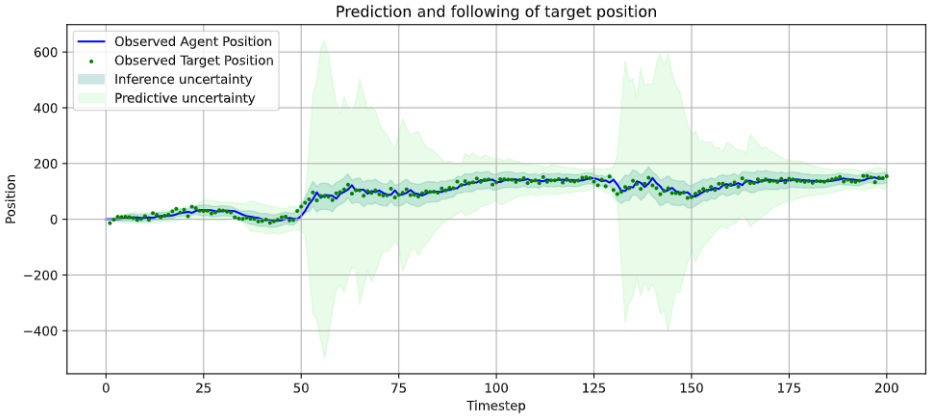


Fig. 2. Relative positions of the agent and the target. Inner shaded area is the inverse precision of the inference of the target position. Outer shaded area is the 68% confidence interval of the predictive posterior

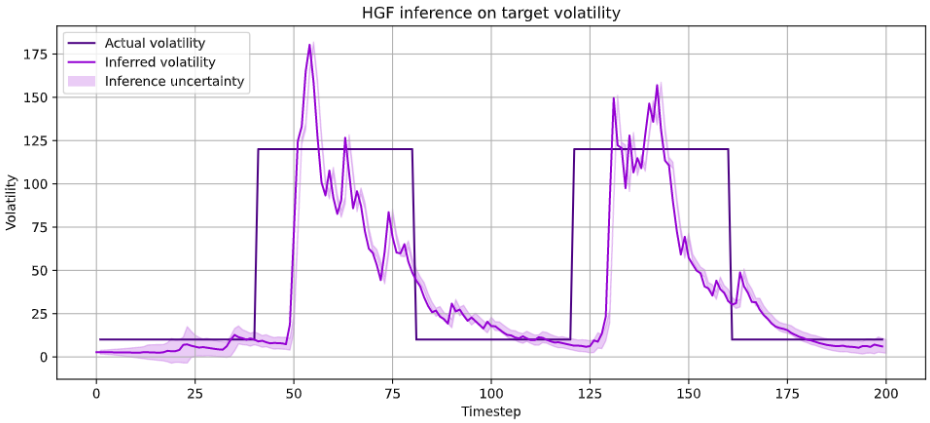


Fig. 3. The agent's HGF-based inference of the volatility (standard deviation) of the target's Gaussian random walk. Shaded area is the inverse precision of the inference.

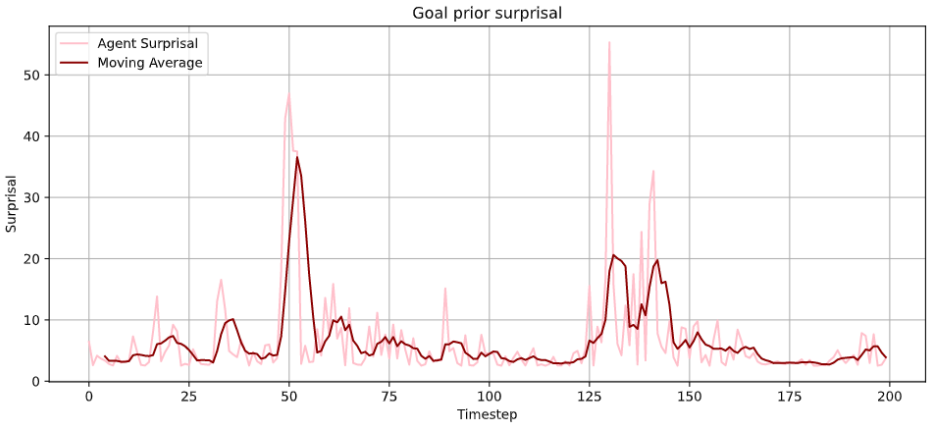


Fig. 4. Agent surprisal at making its observation, as calculated from the goal prior.

5 Discussion

We have here provided a proof-of-principle for active inference models in continuous state space that use hierarchical Gaussian filters to infer and predict the environment. This provides a way of constructing active inference agents that efficiently navigate volatile and noisy continuous state space-environments. The method is consistent with the theoretical framework of active inference, it is elegant, interpretable, and computationally efficient, and shows promise as a method for (re-)extending active inference models into the continuous domain.

There are multiple ways in which this method can be extended in order to employ the active inference framework still more effectively. Most importantly, there is no epistemic component in the current task, which can be included to fully utilize the advantages of active inference. Secondly, we have not here demonstrated the full flexibility of HGF-based active inference. In our simulation, we used a Gaussian goal prior centred around zero, meaning that the agent in practice always moves to the mean of its predictive distribution. Other distributions can be used for the goal prior, however, for example to make the agent prefer observing itself at a certain distance from the target, or be more sensitive to erring in one direction than the other. It is also possible to use more complex instances of the HGF as generative model, allowing for tracking an arbitrary number of possibly inter- or action-dependent hidden states. Since our approach provides a parametric predictive posterior, the expected surprisal can be evaluated directly, leading to the same result as variational methods would converge on. However, when the generative model is more complex, as for example when the agent must plan several steps ahead, or when its actions also influence the movement of the target, this could become less feasible. If so, approximate variational methods might be required.

It would also be valuable to make a more detailed comparison of the HGF-based active inference framework to older continuous state space approaches (for example the saccade models in [3] or the birdsong models in [5]). There are also newer mixed continuous-discrete models which combine discrete policy-level POMDP's with continuous movement and sensation models [9,15,10]. The predictions of the discrete model are then used as prior constraints on the continuous model, which in turn provides evidence for the discrete hypotheses entertained in the former. The main difference between these approaches and HGF-based active inference is that the HGF uses single-step update equations instead of the iterative variational Laplace approach, and that it is generically applicable across contexts. The HGF-based active inference framework can also be contrasted with recent attempts at scaling POMDP methods to complex and continuous domains by amortizing the specification of the generative model with deep learning techniques [19,1,11]. Here, an advantage of HGF-based active inference is that it, beyond specification of hyperparameters as is also the case in deep learning approaches, does not need to be trained, and that it is fully transparent and interpretable.

Finally, it still remains to equip the HGF-based active inference method with parameter and model structure learning capabilities, so that it can select the

HGF-architecture that best explains observations. This is especially important when mapping the HGF to neuronal message passing [20]. Note that this can be combined with recent approaches where the hyperparameters of the HGF are learned online [16,17]. It also remains to apply it more complex environments, and to fit it to experimentally observed behaviour. This is feasible because it has been shown that the HGF is generic and adaptable, can be fit to experimental data, and can be mapped onto neuronal message-passing.

References

1. Çatal, O., Wauthier, S., De Boom, C., Verbelen, T., Dhoedt, B.: Learning generative state space models for active inference. Frontiers in Computational Neuroscience **14**, 103 (2020)
2. Friston, K.: A free energy principle for a particular physics. arXiv preprint arXiv:1906.10184 (2019)
3. Friston, K., Adams, R., Perrinet, L., Breakspear, M.: Perceptions as hypotheses: saccades as experiments. Frontiers in psychology **3**, 151 (2012)
4. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., Pezzulo, G.: Active inference and learning. Neuroscience & Biobehavioral Reviews **68**, 862–879 (Sep 2016). https://doi.org/10.1016/j.neubiorev.2016.06.022
5. Friston, K., Frith, C.: A duet for one. Consciousness and cognition **36**, 390–405 (2015)
6. Friston, K., Kiebel, S.: Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1521), 1211–1221 (2009)
7. Friston, K.J., Daunizeau, J., Kiebel, S.J.: Reinforcement Learning or Active Inference? PLoS ONE **4**(7), e6421 (Jul 2009). https://doi.org/10.1371/journal.pone.0006421
8. Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J.: Action and behavior: A free-energy formulation. Biological Cybernetics **102**(3), 227–260 (Mar 2010). https://doi.org/10.1007/s00422-010-0364-z
9. Friston, K.J., Parr, T., de Vries, B.: The graphical brain: belief propagation and active inference. Network Neuroscience **1**(4), 381–414 (2017)
10. Friston, K.J., Sajid, N., Quiroga-Martinez, D.R., Parr, T., Price, C.J., Holmes, E.: Active listening. Hearing research **399**, 107998 (2021)
11. Van de Maele, T., Verbelen, T., Çatal, O., De Boom, C., Dhoedt, B.: Active vision for robot manipulators using the free energy principle. Frontiers in neurorobotics **15**, 14 (2021)
12. Mathys, C., Daunizeau, J., Friston, K.J., Stephan, K.E.: A Bayesian foundation for individual learning under uncertainty. Frontiers in Human Neuroscience **5**, 39 (2011). https://doi.org/10.3389/fnhum.2011.00039
13. Mathys, C., Lomakina, E.I., Daunizeau, J., Iglesias, S., Brodersen, K.H., Friston, K.J., Stephan, K.E.: Uncertainty in perception and the Hierarchical Gaussian Filter. Frontiers in Human Neuroscience **8**, 825 (2014). https://doi.org/10.3389/fnhum.2014.00825
14. Mathys, C., Weber, L.: Hierarchical gaussian filtering of sufficient statistic time series for active inference. In: International Workshop on Active Inference. pp. 52–58. Springer (2020)

15. Parr, T., Friston, K.J.: The discrete and continuous brain: from decisions to movement—and back again. Neural computation **30**(9), 2319–2347 (2018)
16. Şenöz, İ., De Vries, B.: Online variational message passing in the hierarchical gaussian filter. In: 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2018)
17. Şenöz, İ., de Vries, B.: Online message passing-based inference in the hierarchical gaussian filter. In: 2020 IEEE International Symposium on Information Theory (ISIT). pp. 2676–2681. IEEE (2020)
18. Smith, R., Friston, K., Whyte, C.: A step-by-step tutorial on active inference and its application to empirical data. PsyArXiv (2021)
19. Tschantz, A., Baltieri, M., Seth, A.K., Buckley, C.L.: Scaling active inference. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
20. Weber, Lilian, A.E.: Perception as Hierarchical Bayesian Inference - Toward Non-Invasive Readouts of Exteroceptive and Interoceptive Processing. Doctoral thesis, ETH Zurich (2020)

On Solving a Stochastic Shortest-Path Markov Decision Process as Probabilistic Inference

Mohamed Baioumy, Bruno Lacerda, Paul Duckworth, and Nick Hawes

Oxford Robotics Institute, University of Oxford
{mohamed, bruno, pduckworth, nickh}@robots.ox.ac.uk

Abstract. Previous work on planning as active inference addresses finite horizon problems and solutions valid for *online* planning. We propose solving the general Stochastic Shortest-Path Markov Decision Process (SSP MDP) as probabilistic inference. Furthermore, we discuss online and offline methods for planning under uncertainty. In an SSP MDP, the horizon is *indefinite* and unknown a priori. SSP MDPs generalize finite and infinite horizon MDPs and are widely used in the artificial intelligence community. Additionally, we highlight some of the differences between solving an MDP using dynamic programming approaches widely used in the artificial intelligence community and approaches used in the active inference community. F

1 Introduction

A core problem in the field of artificial intelligence (AI) is building agents capable of automated planning under uncertainty. Problems involving planning under uncertainty are typically formulated as an instance of a Markov Decision Process (MDP). At a high level, an MDP comprises 1) a set of world states, 2) a set of actions, 3) a transition model describing the probability of transitioning to a new state when taking an action in the current state, and 4) an objective function (e.g. minimizing costs over a sequence of time steps). An MDP solution determines the agent's actions at each decision point. An optimal MDP solution is one that optimizes the objective function. These are typically obtained using dynamic programming algorithms [1] [17, 26].

Recent work based on the active inference framework [13] poses the planning problem as a probabilistic inference problem. Several papers have been published showing connections between active inference and dynamic programming to solve an MDP [15, 8, 7]. However, the planning problem being solved in the two communities is not equivalent.

First, dynamic programming approaches used to solve an MDP, such as policy iteration, are valid for finite, infinite and indefinite horizons. Indefinite horizons are finite but of which the length is unknown a priori. For instance, consider

[1] Linear programming approaches are also popular methods for solving MDPs [2, 12, 22, 9]. Additionally, other methods exist in the reinforcement learning community such as policy gradient methods [28, 14, 27].

an agent navigating from a starting state to a goal state in a grid world where the outcome is uncertain (e.g. the 4×4 grid world in Figure 1 shown in the appendix). Before starting to act in the environment, there is no way for the agent to know how many time steps it will take to reach the goal. Algorithms based on dynamic programming, such as policy iteration, are valid for such settings. They can solve the Stochastic Shortest-Path Markov decision process (SSP MDP)[17, 2]. However, work from active inference is only formulated for finite horizons [8].

Second, the optimal solution to an SSP MDP is a stationary deterministic policy [17]. This refers to a mapping from states to actions independent of time. Computing this optimal policy can be done *offline* (without interaction with the environment) or online (while interacting). In the active inference literature however, solving the planning problem is performed by computing a stochastic plan (a sequence of actions given the current state). This is only valid during online planning. Additionally, the solution is only optimal given a certain horizon, which is specified a priori. If the horizon chosen is too short, the agent will not find a solution to reach the target. If it is too long, the solution will be sub-optimal.

The main contribution of this paper is presenting a novel algorithm to solve a Stochastic Shortest-Path Markov Decision Process using probabilistic inference. This is an MDP with an *indefinite horizon*. Additionally, highlighting the several gaps between solving an MDP in the AI community and the active inference community.

Section 2 discusses the SSP MDP and section 3 presets an approach for solving an SSP MDP as probabilistic inference. The equivalence between the two methods is shown in Section 4.1. Furthermore, the difference between world states and temporal state is highlighted in Section 4.2. Policies, plans and probabilistic plans are discussed in Section 4.3. Finally, a discussion on online vs offline planning can be found in Section 5.

2 Stochastic Shortest Path MDP

An SSP MDP is defined as a tuple $\mathcal{M} = (S, A, C, T, G)$. S is the set of states, A is the set of actions, $C : S \times A \times S \to \mathbb{R}$ is the cost function, and $T : S \times A \times S \to [0, 1]$ is the transition function. $G \subset S$ is the set of goal states. Each goal state $s_g \in G$ is absorbing and incurs zero cost. The expected cost of applying action a in state s is $\bar{C}(s, a) = \sum_{s' \in S} T(s, a, s') \cdot C(s, a, s')$. The minimum expected cost at state s is $\bar{C}^*(s) = \min_{a \in A} \bar{C}(s, a)$. A policy maps a state to a distribution over action choices. A policy is deterministic if it chooses a single action at each step. A policy π is proper if it reaches $s_g \in G$ starting from any s with probability 1. In an SSP MDP, the following assumptions are made [17]: a) there exists a proper policy, and b) every improper policy incurs infinite cost at all states where it is improper.

The goal is to find the an *optimal policy* π^* with the minimum expected cost $\bar{C}^*(s)$ and can be computed as

$$\pi^*(s) = \text{argmin}_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}\left(s, a, s'\right) \left[\mathcal{C}\left(s, a, s'\right) + V^*\left(s'\right) \right] \right].$$

$V^*(s)$ is referred to as the optimal value for a state s and is defined as:

$$V^*(s) = \min_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{T}\left(s, a, s'\right) \left[\mathcal{C}\left(s, a, s'\right) + V^*\left(s'\right) \right] \right].$$

Crucially, the optimal policy π^* corresponding to the optimal value function is Markovian (only dependant on the current state) and deterministic [17]. Solving an SSP MDP means finding a policy that minimizes expected cost, as opposed to one that maximizes reward. This difference is purely semantic as the problems are dual. We can define a reward function $\mathcal{R} = -\mathcal{C}$ and move to a reward maximization formulation. A more fundamental distinction is the presence of a special set of (terminal) goal states, in which staying forever incurs no cost.

Solving an SSP MDP can be done using standard dynamic programming algorithms such as policy iteration. Policy iteration can be divided in two steps, policy evaluation and improvement. In policy evaluation, for a policy π, the value function $V_\pi(s)$ is recursively evaluated until convergence as

$$V_\pi(s) \leftarrow \sum_{s' \in \mathcal{S}} \mathcal{T}\left(s, \pi(s), s'\right) \left[\mathcal{C}\left(s, \pi(s), s'\right) + V_\pi\left(s'\right) \right].$$

In the policy improvement step, the state-action value function $Q(s, a)$ is computed as:

$$Q(s, a) = \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') [\mathcal{C}(s, a, s') + V(s')].$$

Then we compute a new policy as $\pi' = \text{argmin}_{a \in \mathcal{A}} Q(s, a)$ for every state in S. Iterating between these two steps guarantees convergence to an optimal policy.

Properties of an SSP MDP. An SSP MDP can be shown to generalize finite, infinite and indefinite horizon MDPs [3, 17]. Thus algorithms valid for an SSP MDP are also valid for the finite and infinite horizon MDPs. Additionally, it can be proven that each SSP MDP has an optimal deterministic policy independent of time. Therefore, the claims made in [8] about active inference being more general since it computes stochastic policies are unjustified when solving an MDP. However, these results do not hold in partially observable cases or in the presence of uncertain models. But in the SSP MDP defined above (which is commonly used in the AI community), there always exists a deterministic optimal policy. Note that there is an infinite number of stochastic policies but only a finite number of deterministic policies ($|\mathcal{S}|^{|\mathcal{A}|}$). This greatly speeds up the algorithms while still guaranteeing optimality.

3 Solving an SSP MDP as probabilistic inference

In this section we discuss a novel approach for solving an SSP MDP as probabilistic inference. We use an inference algorithm that exactly solves an SSP MDP as defined in the previous section. This approach is inspired by work from [32, 31] which solves an MDP with an indefinite horizon. This approach has been successfully applied to solve problems of planning under uncertainty, e.g. [18, 30].

3.1 Definitions

The definition of an SSP MDP includes a set of world states $\mathcal{S}$ and actions $\mathcal{A}$. In probabilistic inference we instead reason about *temporal* states and actions. A temporal state s_t is a random variable defined over all world states. Conceptually it represents the state that the agent will visit at the time-step t. For the grid world in Figure 1, there are 16 world states but the number of temporal states is unknown a priori since the horizon is unknown.

The transition probability is defined as a probability distribution over temporal states and actions as $P(s_{t+1}|a_t, s_t)$. If the random variables are fixed to specific world states i and j and an action a, the transition probability $P(s_{t+1} = j|a_t = a, s_t = i)$ would be equivalent to the transition function $\mathcal{T}$ defined for an SSP MDP. The probability of taking a certain action in a state is parameterized by a policy π as $P(a_t = a|s_t = i; \pi) = \pi_{ai}$. This policy is defined exactly the same as in the case of an SSP MDP.

The cost function $P(c_t|s_t, a_t)$ is defined differently. The temporal cost variables c_t are defined as binary random variables $c_t \in \{0, 1\}$. Translating an arbitrary cost function to temporal costs can be done by scaling the cost function $\mathcal{C}(s, a)$ (as defined in the previous section) between the minimum cost ($\min(\mathcal{C})$) and maximum cost ($\max(\mathcal{C})$) as:

$$P(c_t = 1 \mid a_t = a, s_t = s) = \frac{\mathcal{C}(a, s) - \min(\mathcal{C})}{\max(\mathcal{C}) - \min(\mathcal{C})}.$$

Any expression with $P(c_t = 1)$ can be though of as 'the probability of a cost being maximal'. Thus, the probability of a cost being maximal given a state and an action is $P(c_t = 1 \mid a_t = a, s_t = s)$. Now we can reason about the highest possible cost for a state s and action a as one where $P(c_t = 1 \mid a_t = a, s_t = s) = 1$ and the lowest possible cost as $P(c_t = 1 \mid a_t = a, s_t = s) = 0$. Any other cost will have a probability in-between, according to its magnitude.

Finally, we model the horizon as a random variable. The temporal states and actions are considered up to the end of the horizon T. However, the horizon is generally unknown. We thus model T itself as a random variable. Combining all this information we can define the SSP MDP using a probabilistic model.

3.2 Mixture of finite MDPs

In this section we define the SSP MDP in terms of a mixture of finite MDPs with only a final cost variable. Given every horizon (for instance $T = 1$) the

finite MDP can be given as $P(c, s_{0:T}, a_{0:T} \mid T; \boldsymbol{\pi})$. Note that we dropped the time-index for c_t since there is only one cost variable now. This model can be factorized as

$$P(c, s_{0:T}, a_{0:T} \mid T; \boldsymbol{\pi}) = P(c \mid a_T, s_T)\, P(a_0 \mid s_0; \boldsymbol{\pi})\, P(s_0) \cdot \prod_{t=1}^{T} P(a_t \mid s_t; \boldsymbol{\pi})\, P(s_t \mid a_{t-1}, s_{t-1})$$

To reason about the full MDP, we consider the mixture model of the joint given by the joint probability distribution

$$P(c, s_{0:T}, a_{0:T}, T; \boldsymbol{\pi}) = P(c, s_{0:T}, a_{0:T} \mid T; \boldsymbol{\pi})\, P(T)$$

where $P(T)$ is a prior over the total time, which we choose to be a flat prior (uniform distribution).

3.3 Computing an optimal policy

Our objective is to find a policy that minimizes the expected cost. Similarly to policy iteration, we do not assume any knowledge about the initial state. Expectation-Maximization[2] can be used to find the optimal parameters of our model: the policy π. The E-step will, for a given π, compute a posterior over state-action sequences. The M-step then adapts the model parameters π to optimize the expected likelihood with respect to the quantities calculated in the E-step.

E-step: a backwards pass in all finite MDPs. We use the simpler notation $p(j \mid a, i) \equiv P(s_{t+1} = j \mid a_t = a, s_t = i)$ and $p(j \mid i; \boldsymbol{\pi}) \equiv P(s_{t+1} = j \mid s_t = i; \boldsymbol{\pi}) = \sum_a p(j \mid a, i)\pi_{ai}$. Further, as a 'base' for backward propagation, we define

$$\beta_0(i) = P(c = 1 \mid x_T = i; \boldsymbol{\pi}) = \sum_a P(c = 1 \mid a_T = a, x_T = i)\, \pi_{ai}.$$

This is the immediate cost when following a policy π. It is the expected cost if there is only one time-step remaining. Then, we can recursively compute all the other backward messages. We use the index τ to indicate a backwards counter. This means that $\tau + t = T$, where T is total (unknown) horizon length. This is computed as

$$\beta_\tau(i) = P(c = 1 \mid x_{T-\tau} = i; \boldsymbol{\pi}) = \sum_j p(j \mid i; \boldsymbol{\pi})\beta_{\tau-1}(j).$$

Intuitively, the backward messages are the expected cost if one incurs a cost at the last time step only. So, β_2, is the expected cost if the agent follows the policy π for two time-steps and only incurs a cost after that. Using these messages, we can compute a value function dependent on time, actions and states given as:

[2] An expectation-maximization algorithm can be viewed as performing free-energy minimization [21, 16]. In the E-step, the free-energy is computed and the M-step updates the parameters to minimize the free-energy.

$$q_\tau(a,i) = P\left(c=1 \mid a_t=a, s_t=i, T=t+\tau; \boldsymbol{\pi}\right) = \begin{cases} \sum_j p(j \mid i,a)\beta_{\tau-1}(j) & \tau > 1 \\ P\left(c=1 \mid a_T=a, s_T=i\right) & \tau = 0. \end{cases}$$

Marginalizing out time, we get the state-action value-function

$$P\left(c=1 \mid a_t=a, s_t=i; \boldsymbol{\pi}\right) = \frac{1}{C}\sum_\tau P(T=t+\tau) q_\tau(a,i)$$

where C is a normalization constant. This quantity is the probability of getting a maximum cost given a state and action. It is similar to the $Q(s,a)$ function computed in policy iteration.

M-step: the policy improvement step. The standard M-step in an EM-algorithm maximizes the expected complete log-likelihood with respect to the new parameters π'. Given that the optimal policy for an MDP is deterministic, a greedy M-step can be used. However, our goal is to minimize the log-likelihood in this case as it refers to a the probability of receiving a maximal cost. This can done as

$$\pi' = \underset{a}{\operatorname{argmin}}(P\left(c=1 \mid a_t=a, s_t=i; \boldsymbol{\pi}\right)) \tag{1}$$

This update converges much faster than in a standard M-step. Here an M-step can be used to to obtain a stochastic policy. However, this is unnecessary since the optimal policy is deterministic. Note that there is an infinite number of stochastic policies but a finite number of deterministic ones. In conclusion, a greedy M-step is faster to converge but still guarantees an optimal policy.

4 Connections between the two views

4.1 Exact relationship between policy iteration and planning as probabilistic inference

The messages β computed during backward propagation are exactly equal to the value functions for a single MDP of finite time. The full value function is can therefore be written as the sum of the βs,

$$V_\pi(i) = \sum_T \beta_T(i)$$

since the prior over time $P(T)$ is a uniform prior. If $P(T)$ is not a uniform distribution, this would result in a mixture rather than a sum. The same applies to the relationship between the Q-value function:

$$Q_\pi(a,i) = \sum_T q_T(a,i).$$

Hence, the E-step essentially performs a policy evaluation which yields the classical value function. Given this relation to policy evaluation, the M-step performs an operation exactly equivalent to standard policy improvement. Thus, the EM-algorithm using exact inference is equivalent to Policy Iteration but computes the necessary quantities differently.

One unanswered question is when to stop computing the backward messages. In [32] messages are computed up to a number T_{max}. From this perspective, the planning as inference algorithm presented is equivalent to the so-called *truncated policy iteration* algorithm as opposed to the more common ϵ-*greedy* version.

In the policy evaluation step, one iterates through the state space to update the value $v_\pi(s)$ for every state until a termination criterion is met. An ϵ-greedy criterion means that we stop iterating though the state space once the maximum difference in $V_\pi(s)$ for any s is smaller than a positive small number ϵ. In truncated policy iteration, however, we iterate through the state space $T_{\max}$ times and then perform the policy improvement step. The probabilistic inference algorithm presented in this paper is equivalent to truncated policy iteration if we restrict the maximum number of β messages to be computed.

4.2 World states vs temporal states

In dynamic programming, one reasons over the world states. In the grid world example in Figure 1, this refers to a grid cell. This grid world has 16 world states. In probabilistic inference, one reasons about a *temporal state*. This is a random variable over all world states. The number of temporal states is dependent on how many time-steps the agents acts in the environment (which is often unknown beforehand). An illustration of the difference is given in Fig. 2.

4.3 Policies, plans and probabilistic plans

A classical planning algorithm computes a plan: a sequence of actions. An algorithm like A* or Dijkstra's algorithm [5] can be used to find the optimal path from a stating state to a goal state, given a deterministic world. Crucially, this solution can be computed offline (without interactions with the environment). In a stochastic world, this does not work since the agent can not predict in which states it will end up. However, one can use deterministic planning algorithms for stochastic environments if the path is re-planned online at every time-step. Determinization-based methods have found success in solving planning under uncertainity problems such as the famous FF-replan algorithm [34]

Active inference approaches computes a *probabilistic plan*. The active inference literature calls this a policy; however, we use a different term to avoid confusion.[3] In active inference, the agents computes a finite plan while interacting with the environment. However, rather than assuming a deterministic world

[3] The distinction between a plan and a policy when using active inference has been briefly discussed in [20]. Additionally, other methods computing plans as probabilistic inference have been proposed before active inference in [1, 33]

(like FF-replan [34]), the probabilities are taken into account. This can be shown to compute the optimal solution to an MDP (when planning online). We thus refer to it as a probabilistic plan, a plan that was computed while taking the transition probabilities into account.

Finally, a policy is a mapping from states to actions, i.e. the agent has a preferred action to take for every state. Policies can be stochastic or time-dependent; however, for an SSP MDP the optimal policy is deterministic and independent of time. An agent can compute a policy offline and use it online without needing any additional computation while interacting with the environment. The difference between a plan and policy is illustrated in Fig. 1.

To summarize, a plan or a probabilistic plan can only be used for online planning. Since the outcome of an action is inherently uncertain. Probabilistic plans (as used in active inference) find an optimal solution when used to plan online. A policy also provides an optimal solution and can be computed offline or online.

5 Discussion

In this paper we present a novel approach to solve a stochastic shortest path Markov decision process (SSP MDP) as probabilistic inference. The SSP MDP generalizes many models, including finite and infinite MDPs. Crucially, the dynamic programming algorithms (such as policy iteration) classically used to solve an SSP MDP are valid for *indefinite horizons* (finite but of unknown length); this is not the case for active inference approaches.

The exact connections between solving an MDP using policy iteration and the presented algorithm are discussed. Afterwards, we discussed the gap between solving an MDP in active inference and the approaches in the artificial intelligence community. This included the difference between world states and temporal states, the difference between plans, probabilistic plans and policies. An interesting question now is, which approach is more appropriate? This depends on the problem at hand and whether it can be solved online or offline.

Online and offline planning. As discussed in Section 4.3, a policy is mapping from states to actions and can be used for offline and online planning. Computing a policy is somewhat computationally expensive; however, a look-up is very cheap. Thus if one operates in an environment where the transition and cost function do not change, it is best to compute an optimal policy offline then use it online (while interacting with the environment). This is the case for many planning and scheduling problems, such as a set of elevators operating in sync [6], task-level planning in robotics [19], multi-objective planning [23, 11] and playing games [4, 25]. The challenges in these problems are often that the state-space is incredibly large and thus approximations are needed. However, the problem is fully observable and the cost and transition models are static; the rules of chess do not change half way, for instance.

If the transition or cost functions vary while interacting with the environment (e.g. [24, 10]), an offline solution is not optimal. In this case, the agent can plan online by re-evaluating a policy or computing probabilistic plans (as done in active inference). Computing the latter is cheaper and requires less memory. This is because a probabilistic plan is a distribution over actions $p(a_t)$ up to a time horizon T while a (finite policy) is a conditional distribution $p(a_t|s_t)$ over all world states in S. For any time-step, the posterior over the action is related to the policy such that $p(a_t) = \sum_s p(a_t|s_t)p(s_t)$.

Consider the work in [29, 24]. In both cases a robot operates in an environment susceptible to changes. If the environment changes, the agent can easily construct a new model by varying the cost or transition function but needs to recompute a solution. In [29] the authors recompute a policy at every time-step while in [24] a probabilistic plans is computed using active inference. Since in both cases the solution is recomputed at every time-step, active inference is preferred since it requires less memory and can be computationally cheaper. On the other hand, if the environment only changes occasionally, computing a policy might remain preferable.

To conclude, if the transition and cost functions ($\mathcal{T}$ and $\mathcal{C}$) are static, it is preferable to compute a policy offline. If $\mathcal{T}$ and $\mathcal{C}$ change occasionally, one may still compute an offline policy and recompute a policy only when a change occurs. However, if the environment is dynamic, computing a probabilistic plan (using active inference) is preferable to recomputing a policy at every time-step.

References

1. Attias, H.: Planning by probabilistic inference. In: AISTATS (2003)
2. Bertsekas, D.P., Tsitsiklis, J.N.: An analysis of stochastic shortest path problems. Mathematics of Operations Research **16**(3), 580–595 (1991)
3. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-dynamic programming: an overview. In: Proceedings of 1995 34th IEEE conference on decision and control. vol. 1, pp. 560–564. IEEE (1995)
4. Campbell, M., Hoane, A.J., Hsu, F.: Deep blue. Artif. Intell. **134**, 57–83 (2002)
5. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. MIT press (2009)
6. Crites, R.H., Barto, A.G., et al.: Improving elevator performance using reinforcement learning. Advances in neural information processing systems pp. 1017–1023 (1996)
7. Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., Friston, K.: Active inference on discrete state-spaces: a synthesis. arXiv preprint arXiv:2001.07203 (2020)
8. Da Costa, L., Sajid, N., Parr, T., Friston, K., Smith, R.: The relationship between dynamic programming and active inference: The discrete, finite-horizon case. arXiv preprint arXiv:2009.08111 (2020)
9. d'Epenoux, F.: A probabilistic production and inventory problem. Management Science **10**(1), 98–108 (1963)
10. Duckworth, P., Lacerda, B., Hawes, N.: Time-bounded mission planning in time-varying domains with semi-mdps and gaussian processes (2021)

11. Etessami, K., Kwiatkowska, M., Vardi, M.Y., Yannakakis, M.: Multi-objective model checking of markov decision processes. In: International Conference on Tools and Algorithms for the Construction and Analysis of Systems. pp. 50–65. Springer (2007)
12. Forejt, V., Kwiatkowska, M., Norman, G., Parker, D.: Automated verification techniques for probabilistic systems. In: International school on formal methods for the design of computer, communication and software systems. pp. 53–113. Springer (2011)
13. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: a process theory. Neural computation **29**(1), 1–49 (2017)
14. Grondman, I., Busoniu, L., Lopes, G.A., Babuska, R.: A survey of actor-critic reinforcement learning: Standard and natural policy gradients. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(6), 1291–1307 (2012)
15. Kaplan, R., Friston, K.J.: Planning and navigation as active inference. Biological cybernetics **112**(4), 323–343 (2018)
16. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
17. Kolobov, A.: Planning with Markov Decision Processes: An AI Perspective, vol. 6. Morgan & Claypool Publishers (2012)
18. Kumar, A., Zilberstein, S., Toussaint, M.: Probabilistic inference techniques for scalable multiagent decision making. Journal of Artificial Intelligence Research **53**, 223–270 (2015)
19. Lacerda, B., Faruq, F., Parker, D., Hawes, N.: Probabilistic planning with formal performance guarantees for mobile service robots. The International Journal of Robotics Research **38**(9), 1098–1123 (2019)
20. Millidge, B., Tschantz, A., Seth, A.K., Buckley, C.L.: On the relationship between active inference and control as inference. In: International Workshop on Active Inference. pp. 3–11. Springer (2020)
21. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
22. Nazareth, J.L., Kulkarni, R.B.: Linear programming formulations of markov decision processes. Operations research letters **5**(1), 13–16 (1986)
23. Painter, M., Lacerda, B., Hawes, N.: Convex hull monte-carlo tree-search. In: Proceedings of the International Conference on Automated Planning and Scheduling. vol. 30, pp. 217–225 (2020)
24. Pezzato, C., Hernandez, C., Wisse, M.: Active inference and behavior trees for reactive action planning and execution in robotics. arXiv preprint arXiv:2011.09756 (2020)
25. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815 (2017)
26. Sutton, R.S., Barto, A.G., et al.: Introduction to reinforcement learning, vol. 135. MIT press Cambridge (1998)
27. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems. pp. 1057–1063 (2000)
28. Thomas, P.S., Brunskill, E.: Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines. arXiv preprint arXiv:1706.06643 (2017)

29. Tomy, M., Lacerda, B., Hawes, N., Wyatt, J.L.: Battery charge scheduling in long-life autonomous mobile robots via multi-objective decision making under uncertainty. Robotics and Autonomous Systems **133**, 103629 (2020)
30. Toussaint, M., Charlin, L., Poupart, P.: Hierarchical pomdp controller optimization by likelihood maximization. In: UAI. vol. 24, pp. 562–570 (2008)
31. Toussaint, M., Harmeling, S., Storkey, A.: Probabilistic inference for solving (po) mdps. University of Edinburgh, School of Informatics Research Report EDI-INF-RR-0934 (2006)
32. Toussaint, M., Storkey, A.: Probabilistic inference for solving discrete and continuous state markov decision processes. In: Proceedings of the 23rd international conference on Machine learning. pp. 945–952. ACM (2006)
33. Verma, D., Rao, R.P.: Goal-based imitation as probabilistic inference over graphical models. In: Advances in neural information processing systems. pp. 1393–1400 (2006)
34. Yoon, S.W., Fern, A., Givan, R.: Ff-replan: A baseline for probabilistic planning. In: ICAPS. vol. 7, pp. 352–359 (2007)

A Appendix: Illustrations

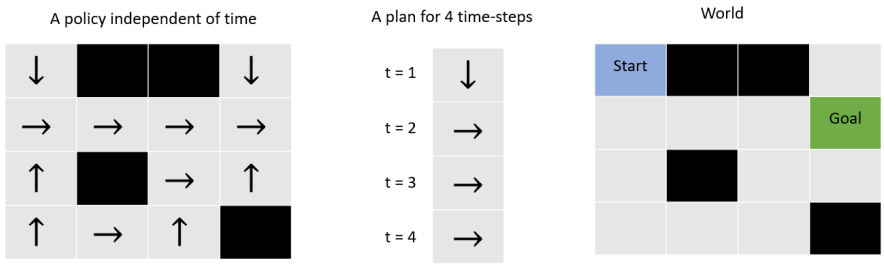


Fig. 1. An illustration of a 4×4 grid world **(right)**. The initial state is blue and goal state is green. An illustration for a policy (**left**) and a plan (**middle**).

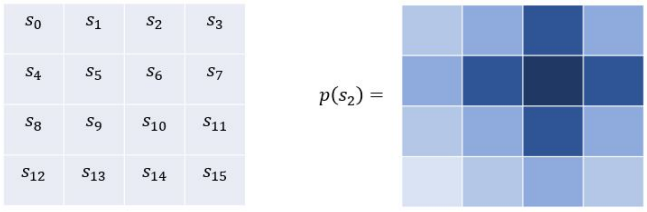


Fig. 2. Annotated world states (**left**) and a posterior over a temporal state (**right**).

Habitual and Reflective Control in Hierarchical Predictive Coding

Paul F Kinghorn[1], Beren Millidge[2] and Christopher L Buckley[3]

[1] School of Engineering and Informatics, University of Sussex p.kinghorn@sussex.ac.uk
[2] MRC Brain Networks Dynamics Unit, University of Oxford beren@millidge.name
[3] School of Engineering and Informatics, University of Sussex c.l.buckley@sussex.ac.uk

Abstract. In cognitive science, behaviour is often separated into two types. Reflexive control is habitual and immediate, whereas reflective is deliberative and time consuming. We examine the argument that Hierarchical Predictive Coding (HPC) can explain both types of behaviour as a continuum operating across a multi-layered network, removing the need for separate circuits in the brain. On this view, "fast" actions may be triggered using only the lower layers of the HPC schema, whereas more deliberative actions need higher layers. We demonstrate that HPC can distribute learning throughout its hierarchy, with higher layers called into use only as required.

Keywords: Hierarchical Predictive Coding · decision making · action selection.

1 Introduction

In the field of cognitive science, behaviour is widely considered to be separated into two classes. Habitual (reflexive) behaviour responds rapidly and instinctively to stimuli, while reflective behaviour involves slower top-down processing and a period of deliberation. These different classes of behaviour have been labelled System 1 and System 2 by Stanovich and West in 2000 [30] and the topic was popularised in Daniel Kahneman's book "Thinking Fast and Slow" in 2011 [21]. However, it has remained unclear how the two processes are implemented in the brain.

This paper investigates how a single system operates when generating behaviours which require different amounts of deliberation. The system does not carry out planning or evaluation of prospective outcomes, but simply compares the triggering of actions which require more or less time to select the correct action for a presented situation. This distinction has some parallels with, but is separate from, the split between goal based planning (often called model-based) and habitual (often called model-free) control [11,31]. Although there is some evidence that separate brain systems underpin goal based planning and habits [18,11,33], there are also indications from fMRI and lesion studies that these processes can co-exist within the same regions of the brain [9,11,33], challenging the notion of separate systems. Since it is possible that even these extremes of behaviour are computed together in the brain, a useful contribution to the topic would be to analyse how reflexive and reflective behaviour can arise from a unified system.

Independent from this debate, there has been much progress in the last decade on the idea that perception and action are both facets of the principle of free energy minimization in the brain [16,14]. According to this approach, the brain maintains a generative model of its environment and uses variational inference to approximate Bayesian inference [29,20]. One way of implementing this (under Gaussian assumptions) is to use a hierarchical predictive coding architecture [28,12,13], which

has successive layers of descending predictions and ascending prediction errors [5,3,32,24]. This theory is also often referred to as Predictive Processing (PP) [7,25]. In this paper, we investigate how PP approaches can explain both reflexive and reflective behaviour simultaneously using a single hierarchical predictive coding network architecture and inference procedure.

In it basic form, PP uses a generative model to try and correctly infer hidden causes for incoming observations. In a hierarchical predictive coding network, all layers of the hierarchy are updated to minimize prediction errors until a fixed point is reached, with the resultant top layer being the best explanation of the hidden causes of the observations at the bottom layer [7,5,3].

PP can also be used to explain action, with the network modelling how actions and sensations interact [2,6,4,8,27,26,25]. Actions are triggered by descending predictions which cause low level prediction errors. These errors are rectified through reflex arcs [17,1,19]. In theory, this means that motor behaviour need not wait for full end-to-end inference to be completed but, rather, action takes place once a threshold has been crossed on the reflex muscle.

This paper investigates the extent to which action selection in a predictive coding network (PCN) relies on all the layers of the PCN. To do this, we train a network to associate actions and observations with each other. We then investigate whether inference across the full network is required in order to trigger the correct action for a given observation. We show that a decision making task with a higher degree of complexity will use more of the layers and may be strongly dependent on the top layer being correctly inferred. Conversely, a decision which is a simple function of sensory observations can operate without involvement of higher layers, despite the fact that learning included those higher layers. This demonstrates that learning allows a hierarchy of action/sensation linkages to be built up in the network, with agents able to use information from lower layers to infer the correct actions without necessarily needing to engage the whole network. These findings suggest that a single PCN architecture could explain both reflexive and reflective behaviour.

In the general case of state space models, the fixed point of a PCN is often in a moving frame of reference. However, the implementation described in this paper ignores state transitions or dynamics and restricts itself to static images. It should therefore not be confused with the notion of predictive coding sometimes seen in the engineering or active inference literature which rests on a state space model for generating timeseries. Rather, our formulation follows the approach of Rao and Ballard's seminal paper [28] and ignores any temporal prediction components, whilst retaining what Friston describes as "the essence of predictive coding, namely any scheme that finds the mode of the recognition density by dynamically minimising prediction error in an input-specific fashion" [12].

The remainder of this paper is set out as follows. Section 2 outlines the HPC model which is used to implement variational inference. Section 3 describes the experiments which we use to analyse inference of labels and actions in PCNs. Section 4 presents the experimental results, demonstrating that learning to act need not rely on high level hidden states. Moreover, we show that the number of higher layers which can be ignored in decision making relates to the complexity of information needed to make that decision.

2 Hierarchical Predictive Coding (HPC)

This section presents a quick overview of how HPC can be used to approximate variational inference. For a more guided derivation, see [5,3,24,16].

At the core of the free-energy principle is the concept that, in order to survive, an agent must strive to make its model of the world a good fit for incoming observations, o. If the model of

observations $p(o)$ can be explained by hidden states of the world s then, in theory, a posterior estimate for s could be obtained using Bayes rule over a set of observations:

$$p(s \mid o) = \frac{p(o \mid s)\ p(s)}{p(o)} = \frac{p(o \mid s)\ p(s)}{\int p(o \mid s)\ p(s)\ ds} \tag{1}$$

but the denominator is likely to be intractable. Therefore s is approximated using variational inference. An auxiliary model (the variational distribution) is created, $q(s;\psi)$, and the divergence between q and the true posterior $p(s \mid o)$ minimized. The KL divergence is used to measure this:

$$KL[q(s;\psi) \parallel p(s \mid o)] = \int q(s;\psi) log \frac{q(s;\psi)}{p(s \mid o)} ds = \mathcal{F} + log\ p(o) \tag{2}$$

where the variational free energy $\mathcal{F}$ is defined as:

$$\mathcal{F} = \int q(s;\psi) log \frac{q(s;\psi)}{p(s,o)} ds \tag{3}$$

The value $-log\ p(o)$, is an information theoretic measure of the unexpectedness of an observation, variously called surprise, suprisal or negative of log model evidence. By adjusting s to minimize surprisal, the model becomes a better fit of the environment. Noting that KL is always positive, it can be seen from equation (2) that $\mathcal{F}$ is an upper bound on surprisal. Therefore, to make the model a good fit for the data, it suffices to minimize $\mathcal{F}$.

The next step is to consider how this would be implemented in the brain via HPC. In HPC, the generative model $p(s,o)$ is implemented in Markovian hierarchical layers, where the priors are simply the values of the layer above, mapped through a weight matrix and a nonlinear function. The prior at the top layer may either be a flat prior or set externally. With N layers, the top layer is labelled as layer 1, and the observation at the bottom as layer N. Thus:

$$p(s,o) = p(s_N \mid s_{N-1}) \ldots p(s_2 \mid s_1)\ p(s_1) \text{ where } s_N = o \tag{4}$$

The generative model is assumed to be Gaussian at each layer,

$$p(s_{n+1} \mid s_n) = N(s_{n+1}; f(\Theta_n\ s_n), \Sigma_{n+1}) \tag{5}$$

where s_n is a vector representing node values on layer n, Θ_n is a matrix giving the connection weights between layer n and layer $n+1$, f is a non-linear function and z_n is Gaussian noise at each layer. Note that the network also has a bias at each layer which is updated in a similar manner to the weights. This has not been included here for brevity. [Here we have shown the form where the argument of f is a weighted linear mixture of hidden states, in order to make clear how we have implemented the hierarchy. But this could equally be generalised to any non linear function f.]

Making the assumption that q is a multivariate Gaussian distribution $q(s) \sim N(s;\mu,\Sigma)$, and further assuming that the distribution of q is tightly packed around μ (to enable use of the Laplace assumption), $\mathcal{F}$ reduces to:

$$\mathcal{F} \approx -log\ p(\mu, o) \tag{6}$$

where o is a vector representing observations and μ is the mean of the brain's probability distribution for s. It is important to note that in this paper the observations are not confined to incoming senses but also include actions, in the form of proprioceptive feedback. Exteroceptive observations cause

updates to model beliefs which, in turn, result in updated beliefs on proprioceptive observations. These drive motoneurons to eliminate any prediction error through reflex arcs [17,15,29]. Action can therefore be thought of as just a particular type of observation.

Using the distribution for a multivariate Gaussian, the estimate of $\mathcal{F}$ can be transformed into:

$$\mathcal{F} \approx -log\ p(\mu, o) = \sum_n log\ p(\mu_{n+1} \mid \mu_n) = \sum_n -\frac{1}{2}\epsilon_{n+1}^T \Sigma_{n+1}^{-1} \epsilon_{n+1} - \frac{1}{2} log(2\pi\ |\Sigma_{n+1}|) \tag{7}$$

where $\epsilon_{n+1} := \mu_{n+1} - f(\Theta_n\ \mu_n)$ is the difference between value of layer $n+1$ and the value predicted by layer n. $\mathcal{F}$ is then minimized following the Expectation-Minimization approach [10,23], by using gradient descent to alternately update node values (μ) on a fast timescale and weight values (Θ) on a slower timescale.

The gradient for node updates in a hidden layer uses the values of ϵ_n and ϵ_{n+1}, and is given by the partial derivative:

$$\frac{\partial \mathcal{F}}{\partial \mu_n} = \epsilon_{n+1}\ \Sigma_{n+1}^{-1}\ \Theta_n^T\ f'(\Theta_n\ \mu_n) - \epsilon_n\ \Sigma_n^{-1} \tag{8}$$

but if the node values of the top layer are being updated then this is truncated to only use the difference compared to the layer below:

$$\frac{\partial \mathcal{F}}{\partial \mu_1} = \epsilon_2\ \Sigma_2^{-1}\ \Theta_1^T\ f'(\Theta_1\ \mu_1) \tag{9}$$

As pointed out earlier, downward predictions not only predict exteroceptive (sensory) signals, but also create a proprioceptive prediction error in the motor system (which is cancelled by movement via a reflex arc). In this paper we simply intend to monitor the signals being sent to the motor system and do not wish to include the error cancellation signal being fed back from the reflex arc. For this reason, the update of the "observation node" in the motor system is shown as only using the difference to the layer above:

$$\frac{\partial \mathcal{F}}{\partial \mu_N} = -\epsilon_N\ \Sigma_N^{-1} \tag{10}$$

After the node values have been changed, $\mathcal{F}$ is then further minimized by updating the weights using:

$$\frac{\partial \mathcal{F}}{\partial \Theta_n} = \epsilon_{n+1}\ \Sigma_{n+1}^{-1}\ \mu_n^T\ f'(\Theta_n\ \mu_n) \tag{11}$$

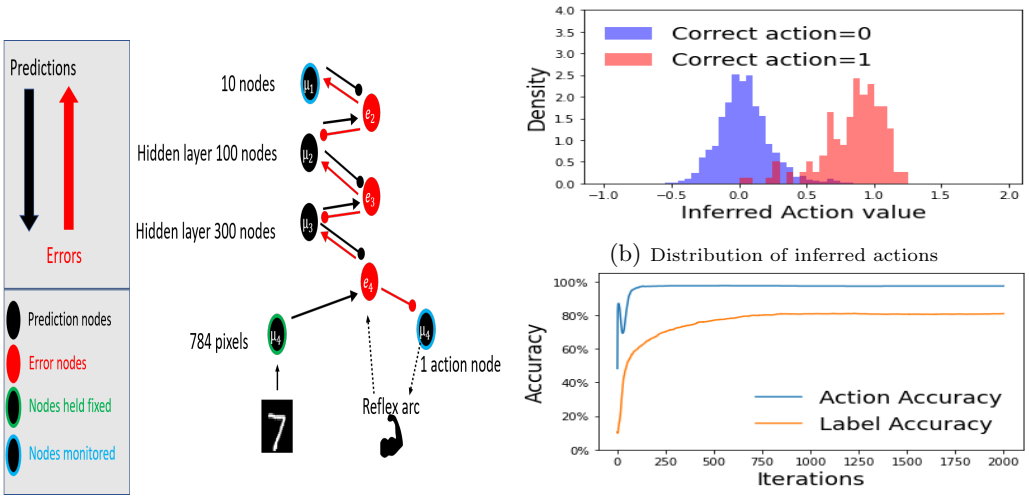
Since the impact of variance is not the primary focus here, our simulations assume that all Σ have fixed values of the identity matrix and therefore the gradient update for Σ has not been included.

Fig. 1a summarises the flow of information in the network during gradient descent update of node values.

3 Methods

Three sets of experiments were designed, to investigate how the process of inference is distributed through hierarchical layers. The first two experiments were each run on three different tasks. The third experiment was run on a single task. The experiments are described below.

In the first set of experiments, we trained three PCNs to carry out separate inference tasks, based on selecting the correct action for a given MNIST image [22]. In all three networks, the observation layer at the bottom of the network contains 785 nodes, made up of 784 sensory nodes (representing the pixels of an MNIST image) and a single binary action node. The top layer uses a one-hot representation of each of the possible MNIST labels. There are two hidden layers of size 100 and 300. Thus, if there are 10 possible labels, there is a four-layer network of size [10,100,300,785], whose generative model produces an MNIST image and an action value from a given MNIST label. The role of each of the networks is, on presentation of an MNIST image, to infer the correct MNIST label at the top and the correct action associated with that image (Fig. 1a).



(a) Network in test mode

(b) Distribution of inferred actions

(c) Development of label and action accuracy

Fig. 1: Results for MNIST-digit1: correct action=1 if label=1, or 0 otherwise. (a) Test configuration. The observation nodes representing the MNIST picture are held fixed and both the label and the action nodes are updated using variational inference. (b) Distribution of inferred actions at end of inference period. (c) Simultaneous development of accuracy for label and action as inference progresses through iterations. Using argmax on label nodes, label accuracy = 80.8%. Using a heaviside function centred on 0.5, action accuracy = 97.4%. This indicates that the correct label is not required in order to select the correct action.

We investigated the relationship between the accuracies of action inference and label inference. Specifically, we asked: to what extent can the action be correctly triggered without correct label inference?

In the first task, MNIST-digit1, we trained the action node to output value 1 if the presented MNIST image has label 1, and value 0 for all other digits, i.e. the job of the action node is to fire when an image of the digit 1 is presented. The network is trained in a supervised manner to learn the generative model, by fixing the top and bottom layers with the training labels and observations

respectively, and then, minimizing $\mathcal{F}$ in an expectation–maximization (EM) fashion [10,23], as described in Section 2. Once trained, the network is then tested for its ability to infer the correct label and action for a given image. This is done by presenting an MNIST image to the 784 sensory states and allowing both the labels at the top and the action at the bottom to update via the variational inference process, according to equations (8) - (10). Updates to the network are applied over a large number of iterations and, at any stage of this process, the current inferred label can be read out as the argmax of the top-layer nodes while the selected action is read out according to a heaviside function applied to the action node value, centred on 0.5.

In the second task, MNIST-groups, we trained the action node to fire if the MNIST label is less than 5, and not fire otherwise. This network is trained and tested using the same process as above.

In the third task, MNIST-barred, half of the MNIST images had a white horizontal bar applied across the middle of the image. A new set of labels was created so that there were now 20 possible labels - 0 to 9 representing the digits without bars, and 10 to 19 representing the digits with bars. Action value 1 was associated with labels 10 to 19, and action value 0 with labels 0 to 9. The network for this task has size [20,100,300,785]. It is trained and tested as for the first two tasks. Appendix A gives full details of the hyperparameters used in the three PCNs.

The second set of experiments used the same three tasks but, instead of fixing MNIST labels to the top of the network in training, the top layer was populated with random noise. The purpose of these experiments was to determine whether the provision of label information in training had any impact on the network's ability to infer the correct action. We then ablated layers from the PCNs in order to investigate the contribution which each layer makes towards inferring the correct action.

The third experiment trained a network where both the MNIST image and the MNIST one hot-labels were placed at the bottom. Above this were 6 layers, all initialized with noisy values. The top layer was allowed to vary freely (see Fig. 4a). This was used to investigate how label inference performed in this scenario (rather than the traditional case of label at the top and image at the bottom), and how performance reacted to ablation of layers in test mode.

4 Results

We first investigated the relationship between accuracy of action and label inference for MNIST-digit1 (where the action node should fire if the MNIST label is 1). When run on a test set of images, the network generates values on the action node which correctly split into two groups centred near to 0 and 1, with a small overlap (Fig. 1b). As a result, the network is able to correctly infer the action for a presented image in over 97% of cases. On the other hand, the label is only correctly inferred in 81% of cases, demonstrating that action selection does not depend entirely on correct label inference. Fig. 1c presents the development of label and action accuracies as iterations progress, confirming that a) action accuracy is always better than label accuracy, b) further iterations will not change this and c) action inference reaches asymptotic performance quicker than label inference.

Fig. 2 compares label and action accuracy for all three tasks. In the MNIST-group task, action accuracy appears to be constrained by label accuracy. In the MNIST-barred task, the correct action is always inferred, even though the network has relatively poor label accuracy. It would therefore seem that the MNIST-group task is reliant on upper layer values in order to select the correct action, whereas the simpler tasks can reach, or approach, optimal action performance regardless of the upper layer values.

However, it is not clear from these results whether the MNIST-group task is relying on the fact that the higher layers contain information about the image labels (recall that this is how the network was trained) or whether it is simply that the existence of the higher layers is providing more compute power. To investigate this, the second set of experiments were run, where the three networks are trained with random noise at the top layer instead of the image label. In testing, label accuracy was now no better than random (as one would expect), but action accuracy was indistinguishable from the original results of Fig. 2. This demonstrates that it is the existence of the layers, rather than provision of label information in training which is driving action inference.

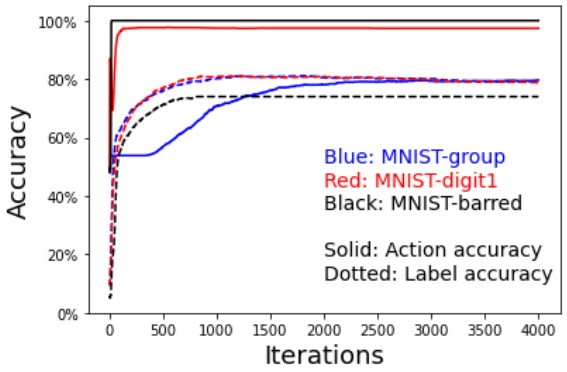
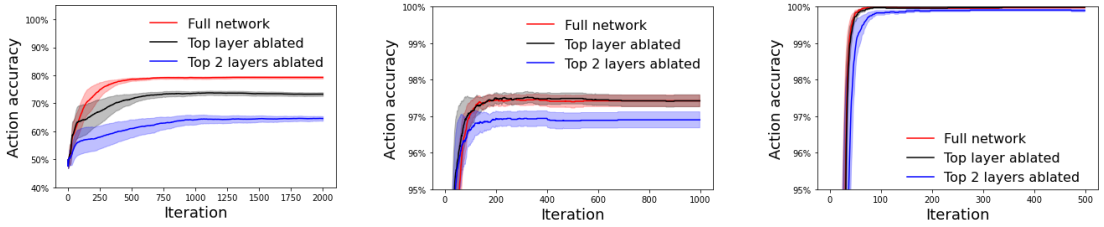


Fig. 2: Development of inference accuracy for label (at top of network) and action (at bottom of network), for 3 different tasks. Blue lines: MNIST-group. Action accuracy is constrained by label accuracy at around 80%. Red lines: MNIST-digit1. Action selection is above 97% accuracy, with label accuracy around 80%. Black lines: MNIST-barred. Accuracy of action selection quickly reaches 100% despite label accuracy being low. On a poorly trained network the results are even more striking, with action accuracy still perfect despite label accuracy of only 20% (results not shown).

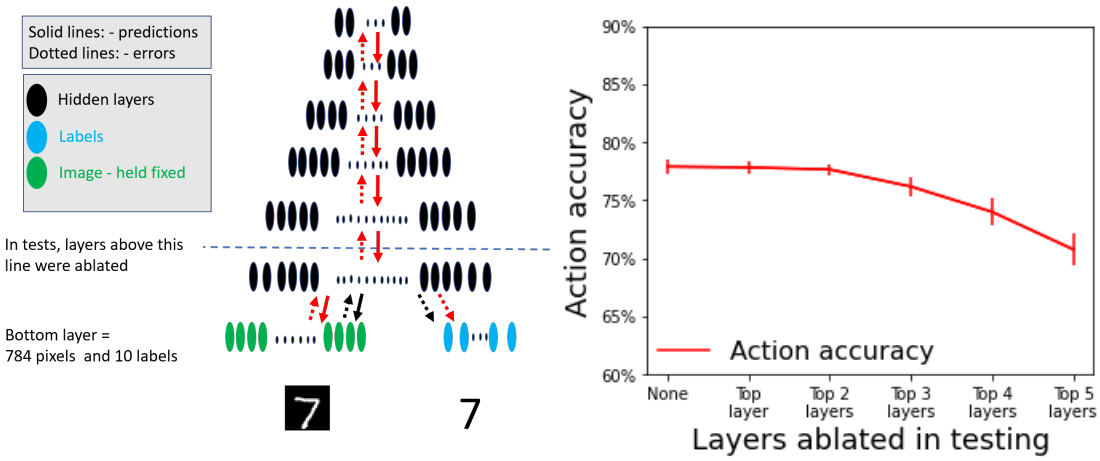
To confirm that the three tasks make different use of the higher layers, action accuracy was measured when the top two layers were ablated in test mode (they were still present in training). Performance on the MNIST-group task (Fig. 3a) deteriorates significantly as the layers are ablated. Conversely, ablation of the top layer has no impact on the action accuracy of either the MNIST-barred (Fig. 3c) or MNIST-digit1 tasks (Fig. 3b). Both suffer slightly if the top 2 layers are ablated, although in the case of MNIST-barred the accuracy only moves from 100% to 99.9%. It can be concluded from these ablation experiments that reliance on higher layers varies with the nature of the task. Tasks which are more challenging may rely on the higher layers, while simple tasks may not suffer at all if the layers are ablated - presumably because all the information required for action selection is entirely available in the lower layers.



(a) MNIST-group (b) MNIST-digit1 (c) MNIST-barred

Fig. 3: Effect of ablating layers on action accuracy. The three tasks cope differently with ablation of layers, as shown in (a), (b) and (c). Note that different y-scales are used on the figures for clarity. Each network was trained using 6 different seeds, and error bars show standard error. Results suggest that, if the lower layers are sufficient for action selection then the higher layers can be ignored.

In the third experiment, we constructed a network with both MNIST image and MNIST one hot-labels at the bottom, representing 10 different binary actions to select from (see Fig. 4a). Above this were 6 layers, all initialized with noisy values (details in Appendix A). Training was carried out as before, presenting a set of images and labels at the bottom of the network and leaving the network to learn weights throughout the hierarchy. The effect of layer ablation on the ability of the network to select the correct action (which in this experiment is the one-hot label) was then tested. When using all the layers, this network produces comparable results to the more standard PCN setup with label at the top and image at the bottom.[4] Ablation results are shown in Fig. 4b. These are consistent with the previous experiments, with accuracy reducing (but still much better than chance value of 10%) as the layers are ablated. In this case it would appear that the top 2 layers are adding nothing to the network's action selection ability. A key point to note is that the learning of the weights was not dependent on the provision of any information at the top of the network - all the learning comes about as a result of information presented at the bottom. Despite this, the network has distributed its ability through several layers, with the major part of successful inference relying on information towards the bottom of the network.



(a) Schematic of network

(b) Effect of layer ablation

Fig. 4: A 7 layer PCN where 10 binary actions are associated with MNIST images. (a) Image and one-hot labels both at the bottom. For ease of reading, the nodes shown on each layer represent both value and error nodes. Red lines show flow of information with no ablation. Black line shows flow if 5 layers are ablated. (b) Ablation of top two layers has no effect on accuracy of action selection. Ablation of the next 3 layers steadily reduces accuracy. Error bars are standard deviations across 10 differently seeded networks.

5 Discussion

We have demonstrated that, when training a PCN with senses and actions at the bottom layer, it is not necessary to provide a high level "hidden state" in training in order to learn the correct actions

[4] At approximately 78%, the accuracy we achieved is significantly lower than standard non-PCN deep learning methods. This is partly because the model has not been fine-tuned (e.g. hyper-parameters, using convolutional layers, etc). But it is also true that generative models tend to underperform discriminative models in classification tasks. This will be particularly true in our implementation which uses flat priors.

for an incoming sensation. Furthermore, the network appears to distribute its learning throughout the layers, with higher layers called into use only as required. In our experiments, this meant that higher layers could be ignored if the lower layers alone contained sufficient information to select the correct action. In effect, the network has learned a sensorimotor shortcut to select the correct actions. On the other hand, if the higher layers contain information which improves action selection, then ablation of those layers reduces, but doesn't destroy, performance - ablation leads to graceful degradation. This flexibility is inherent in the nature of PCNs, unlike feed forward networks, which operate end to end.

Importantly, this suggests that a PCN framework can help explain the development of fast reaction to a stimulus, even though the learning process involves all layers. For example, driving a car on an empty road might only require involvement of lower layers, whereas heavy traffic or icy conditions would require higher layers to deal with the more complex task. The fact that simple short-cuts can arise automatically during training and that the agent can dynamically select actions without involvement of higher layers could possibly also help explain why well-learned tasks can be carried out without conscious perception.

While we have provided an illustrative 'proof of principle' of this approach, much more can be done to investigate how this leads to a continuum of behaviour in active agents, which we list below in no particular order. Firstly, in our experiments inference took place with no influence from above and we have not considered the impact which exogenous priors would have. Secondly, we included no concept of a causal link between action and the subsequent sensory state. Action in real-life situations is a rolling process, with actions impacting subsequent decisions. Because our generative model did not consider time or state transitions, we cannot generalise to active inference in the sense of planning. One might argue that policy selection in active inference is a better metaphor for reflective behaviour, leading to a distinction between reflexive 'homeostatic' responses and more deliberative 'allostatic' plans. Having said this, it seems likely that the same conclusions will emerge. In other words, the same hierarchical generative model can explain reflective and reflexive behaviour at different hierarchical levels. Thirdly, the role of precisions has not been examined. Updating precisions should allow investigation of the role of attention. Finally, we have assumed the existence of a well trained network, and only touched on the performance of a partially trained network. It would be instructive to investigate how reliance on higher layers changes during the learning process.

These results support the view that a predictive coding network in the brain does not need to work from end to end, and can restrict itself to the number of lower layers required for the task at hand, possibly only in the sensorimotor system. There is the possibility of some tentative links here with more enactivist theories of the brain which posit that "representations" encode predicted action opportunities, rather than specify an abstract state of the world, but much further analysis is needed to investigate possible overlaps.

6 Acknowledgements

PK would like to thank Alec Tschantz for sharing the "Predictive Coding in Python" codebase `https://github.com/alec-tschantz/pypc` on which the experimental code was based. Thanks also to three anonymous reviewers whose comments helped improve the clarity of this paper, particularly in relation to temporal aspects of predictive coding. PK is funded by the Sussex Neuroscience 4-year PhD Programme. CLB is supported by BBRSC grant number BB/P022197/1.

References

1. Adams, R.A., Shipp, S., Friston, K.J.: Predictions not commands: active inference in the motor system. Brain Structure and Function **218**(3), 611–643 (2013)
2. Baltieri, M., Buckley, C.L.: Generative models as parsimonious descriptions of sensorimotor loops. The Behavioral and brain sciences **42**, e218–e218 (2019)
3. Bogacz, R.: A tutorial on the free-energy framework for modelling perception and learning. Journal of mathematical psychology **76**, 198–211 (2017)
4. Bruineberg, J., Kiverstein, J., Rietveld, E.: The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. Synthese (Dordrecht) **195**(6), 2417–2444 (2018)
5. Buckley, C.L., Chang, S.K., McGregor, S., Seth, A.K.: The free energy principle for action and perception: A mathematical review (2017)
6. Burr, C.: Embodied decisions and the predictive brain. In: Wiese, T.M..W. (ed.) Philosophy and predictive processing. MIND Group, Frankfurt am Main (2016)
7. Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. The Behavioral and brain sciences **36**(3), 181–204 (2013)
8. Clark, A.: Predicting peace: The end of the representation wars. Open MIND. Frankfurt am Main: MIND Group (2015)
9. Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J.: Model-based influences on humans' choices and striatal prediction errors. Neuron **69**(6), 1204–1215 (2011)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
11. Dolan, R., Dayan, P.: Goals and habits in the brain. Neuron (Cambridge, Mass.) **80**(2), 312–325 (2013)
12. Friston, K.: Learning and inference in the brain. Neural Networks **16**(9), 1325–1352 (2003)
13. Friston, K.: A theory of cortical responses. Philosophical transactions of the Royal Society B: Biological sciences **360**(1456), 815–836 (2005)
14. Friston, K.: The free-energy principle: a unified brain theory? Nature reviews. Neuroscience **11**(2), 127–138 (2010)
15. Friston, K.: What is optimal about motor control? Neuron **72**(3), 488–498 (2011)
16. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. Journal of physiology-Paris **100**(1-3), 70–87 (2006)
17. Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J.: Action and behavior: a free-energy formulation. Biological cybernetics **102**(3), 227–260 (2010)
18. Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P.: States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron **66**(4), 585–595 (2010)
19. Hipólito, I., Baltieri, M., Friston, K., Ramstead, M.J.: Embodied skillful performance: Where the action is. Synthese pp. 1–25 (2021)
20. Hohwy, J.: The predictive mind. Oxford University Press (2013)
21. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
22. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010), `http://yann.lecun.com/exdb/mnist/`
23. MacKay, D.J., Mac Kay, D.J.: Information theory, inference and learning algorithms. Cambridge university press (2003)
24. Millidge, B.: Combining active inference and hierarchical predictive coding: A tutorial introduction and case study. PsyArXiv (2019)
25. Pezzulo, G., Donnarumma, F., Iodice, P., Maisto, D., Stoianov, I.: Model-based approaches to active perception and control. Entropy (Basel, Switzerland) **19**(6), 266 (2017)
26. Pezzulo, G., Rigoli, F., Friston, K.: Active inference, homeostatic regulation and adaptive behavioural control. Progress in neurobiology **134**, 17–35 (2015)

27. Ramstead, M.J., Kirchhoff, M.D., Friston, K.J.: A tale of two densities: active inference is enactive inference. Adaptive behavior **28**(4), 225–239 (2020)
28. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience **2**(1), 79–87 (1999)
29. Seth, A.K.: The cybernetic bayesian brain: from interoceptive inference to sensorimotor contingencies (2015)
30. Stanovich, K.E., West, R.F.: Individual differences in reasoning: Implications for the rationality debate? Behavioral and brain sciences **23**(5), 645–665 (2000)
31. Sutton, R.S.: Reinforcement learning : an introduction (2018)
32. Whittington, J.C., Bogacz, R.: An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. Neural computation **29**(5), 1229–1262 (2017)
33. Wunderlich, K., Dayan, P., Dolan, R.J.: Mapping value based planning and extensively trained choice in the human brain. Nature neuroscience **15**(5), 786–791 (2012)

A Network parameters

Network size: 4 layer
Number of nodes on each layer: 10, 100, 300, 785 for MNIST-group and MNIST-digit1. 20, 100, 300, 785 for MNIST-barred. In the bottom layer, 784 nodes were fixed to the MNIST image, the 785^{th} node was an action node which updates in testing. In initial set of experiments, top layer was fixed to a one-hot representation of MNIST label in training. In second set of experiments this was set to random value and allowed to update.
Non-linear function: tanh
Bias used: yes
Training set size: full MNIST training set of 60,000 images, in batches of 640
Number of training epochs: 10
Testing set size: 1280 images selected randomly from MNIST test set
Learning parameters used in weight update of EM process: Learning Rate= 1e-4, Adam
Learning parameters used in node update of EM process: Learning Rate= 0.025, SGD
Number of SGD iterations in training: 200
Number of SGD iterations in test mode: 200 * epoch number. The size is increased as epochs progress to allow for the decreasing size of the error between layers (as discussed in the text, this would normally be counteracted by increase in precision values).
Random initialisation: Except where fixed, all nodes were initialized with a random values selected from $\mathcal{N}(0.5, 0.05)$

In the experiment using a 7 layer network, the number of nodes on each layer were: 10, 25, 50, 100, 200, 300, 794. All other parameters the same as above

Deep Active Inference for Pixel-Based Discrete Control: Evaluation on the Car Racing Problem

N.T.A. van Hoeffelen and Pablo Lanillos

Department of Artificial intelligence
Donders Institute for Brain, Cognition, and Behaviour
Radboud University
Montessorilaan 3, 6525HR Nijmegen, the Netherlands
niels.vanhoeffelen@ru.nl
p.lanillos@donders.ru.nl

Abstract. Despite the potential of active inference for visual-based control, learning the model and the preferences (priors) while interacting with the environment is challenging. Here, we study the performance of a deep active inference (dAIF) agent on OpenAI's car racing benchmark, where there is no access to the car's state. The agent learns to encode the world's state from high-dimensional input through unsupervised representation learning. State inference and control are learned end-to-end by optimizing the expected free energy. Results show that our model achieves comparable performance to deep Q-learning. However, vanilla dAIF does not reach state-of-the-art performance compared to other world model approaches. Hence, we discuss the current model implementation's limitations and potential architectures to overcome them.

Keywords: Deep Active Inference · Deep Learning · POMDP · Visual-based Control

1 Introduction

Learning from scratch which actions are relevant to succeed in a task using only high-dimensional visual input is challenging and essential for artificial agents and robotics. Reinforcement learning (RL) [26] is currently leading the advances in pixel-based control, e.g., the agent learns an action policy that maximizes the accumulated discounted rewards. Despite its dopamine biological inspiration, RL is far from capturing the physical processes happening in the brain. We argue, that prediction in any form (e.g., visual input, muscle feedback or dopamine) may be the driven motif of the general learning process of the brain [14]. Active inference [8,4], a general framework for perception, action and learning, proposes that the brain uses hierarchical generative models to predict incoming sensory data [6] and tries to minimize the difference between the predicted and observed sensory signals. This difference, mathematically described as the variational free energy (VFE), needs to be minimized to generate better predictions about the world that causes these sensory signals [8,17]. Through acting on its environment,

an agent can affect sensory signals to be more in line with predicted signals, which in turn leads to a decrease of the error between observed and predicted sensory signals.

AIF models have shown great potential in low-dimensional and discrete state spaces. To work in higher-dimensional state spaces, deep active inference (dAIF) has been proposed, which uses deep neural networks for approximating probability density functions (e.g., amortised inference) [17,27,29,2,23]. Interestingly, dAIF can be classified as a generalization of the world models approach [10] and can incorporate reward-based learning, allowing for a direct comparison to RL methods. Since the first attempt of dAIF [29], developments have happened concurrently in adaptive control [23,16] and in planning, exploiting discrete-time optimization, e.g., using the expected free energy [21,28]. In [17], a dIAF agent was tested on several environments in which the state is observable (Cartpole, Acrobot, Lunar-lander). In [5], a dAIF agent using Monte-Carlo sampling was tested on the Animal-AI environment. In [3], dAIF tackled the mountain car problem and was also tested on OpenAI's car racing environment. For the car racing environment, they trained the dAIF agent on a handful of demonstration rollouts and compared it to a DQN that interacted with the environment itself. Their results showed that DQN needs a lot more interaction with the environment to start obtaining rewards compared to dAIF trained on human demonstrations of the task. Relevant for this work, in [11], a dAIF agent solved the Cartpole environment as both MDP and as POMPD instances, training the agent on just visual input.

In this paper, we study a dAIF agent[1] based on the proposed architectures in [17,11] for a more complex pixel-based control POMDP task, namely the OpenAI's Car Racing environment [13], and discuss its advantages and limitations compared to other state-of-the-art models. The performance of the dAIF agent was shown to be in line with previous works and on-par with deep Q-learning. However, it did not achieve the performance of other world model approaches [1]. Hence, we discuss the reasons for this, as well as architectures that may help to overcome these limitations.

2 Deep Active Inference Model

The dAIF architecture studied is based on [11] and described in Fig. 1. It makes use of five networks to approximate the densities of Eq. (2): observation (encoding and decoding), transition, policy, and value. The full parameter description of the networks can be found in the Appendix A.

Variational Free Energy (VFE). AIF agents infer their actions by minimizing the VFE expressed at instant k (with explicit action 1-step ahead) as:

$$\mathcal{F}_k = -\mathbf{KL}[q(s_k, a_k) \mid\mid p(o_k, s_{0:k}, a_{0:k})] \tag{1}$$

[1] The code can be found at https://github.com/NTAvanHoeffelen/DAIF_CarRacing

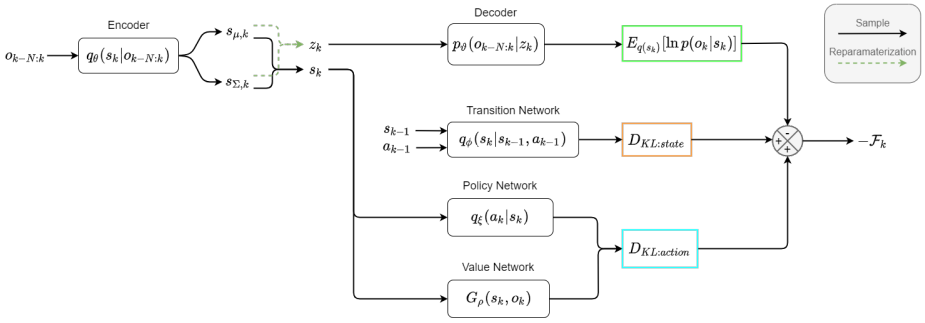


Fig. 1: Deep Active Inference architecture. Five deep artificial neural networks are used to model the observation encoding and decoding, the state transition, the policy and the EFE values. The architecture is trained end-to-end by optimizing the expected variational free energy.

Where s_k, o_k, a_k are the state, the observation and action respectively, $q(s_k, a_k)$ is the recognition density, and $p(o_k, s_{0:k}, a_{0:k})$ the generative model. Under the Markov assumption and factorizing [17,11], Eq. (1) can be rewritten as:

$$\mathcal{F}_k = \underbrace{E_{q(s_k)}[\ln p(o_k|s_k)]}_{\text{sensory prediction}} - \underbrace{\mathbf{KL}[q(s_k) \mid\mid p(s_k|s_{k-1}, a_{k-1})]}_{\text{state prediction}} - \underbrace{\mathbf{KL}[q(a_k|s_k) \mid\mid p(a_k|s_k)]}_{\text{action prediction}} \tag{2}$$

Sensory prediction. The observation network predicts the observations—first term in Eq. (2)—and encodes the high-dimensional input to states $q_\theta(s_k|o_{k-N:k})$ and decodes latent spaces to observations $p_\vartheta(o_{k-N:k}|z_k)$. It can be implemented with a variational autoencoder, where latent representation is described as a multivariate Gaussian distribution with mean s_μ and variance s_Σ. The latent space z_k is obtained using the reparametrisation trick. To train the network, we use the binary cross-entropy and the KL regularizer to force the latent space to be Gaussian:

$$\begin{aligned} L_{o,k} &= -E_{q_\theta(s_k|o_{k-N:k})}[\ln p_\vartheta(o_{k-N:k}|z_k)] \\ &= BCE(\hat{o}_{k-N:k}, o_{k-N:k}) - \frac{1}{2}\sum(1 + \ln s_{\Sigma,k} - s_{\mu,k}^2 - s_{\Sigma,k}) \end{aligned} \tag{3}$$

State prediction. The transition network models a distribution that allows the agent to predict the state at time k given the state and action at time $k-1$, where the input state consists of both the mean s_μ and variance s_Σ. Under the AIF approach this is the difference between the state distribution (generated by the transition network) and the actual observed state (from the encoder): $\mathbf{KL}[q(s_k) \mid\mid p(s_k|s_{k-1}, a_{k-1})]$. For the sake of simplicity, we define a

feed-forward network that computes the maximum-a-posteriori estimate of the predicted state $\hat{s}_k = q_\phi(s_{k-1}, a_{k-1})$ and train it using the mean squared error:

$$MSE(s_{\mu,k}, q_\phi(s_{k-1}, a_{k-1})) \tag{4}$$

Action prediction. We use two networks to evaluate action: the policy and value network. The action prediction part of Eq. (2) is the difference between the model's action distribution and the "optimal" true distribution. It is a KL divergence, which can be split into an energy and an entropy term:

$$\begin{aligned}\mathbf{KL}[q(a_k|s_k) \,||\, p(a_k|s_k)] &= -\sum_a q(a_k|s_k) \ln \frac{p(a_k|s_k)}{q(a_k|s_k)} \\ &= \underbrace{-\sum_a q(a_k|s_k) \ln p(a_k|s_k)}_{\text{energy}} \underbrace{-\sum_a q(a_k|s_k) \ln q(a_k|s_k)}_{\text{entropy}}\end{aligned} \tag{5}$$

The policy network models the distribution over actions at time k given the state at time k $q_\xi(a_k|s_k)$. It is implemented as a feed-forward neural network that returns a distribution over actions given a state.

The value network computes the Expected Free Energy (EFE) [21,17,11] which is used to model the true action posterior $p(a_k|s_k)$. As the true action posterior is not exactly known, we assume that prior belief makes the agent select policies that minimize the EFE. We model the distribution over actions as a precision-weighted Boltzmann distribution over the EFE [21,17,5,24]:

$$p(a_k|s_k) = \sigma(-\gamma G(s_{k:N}, o_{k:N})) \tag{6}$$

where $G(s_{k:N}, o_{k:N})$ is the EFE for a set of states and observations up to some future time N. As we are dealing with discrete time steps, it can be written as a sum over these time steps:

$$G(s_{k:N}, o_{k:N}) = \sum_k^N G(s_k, o_k) \tag{7}$$

The EFE is evaluated for every action, because of this we implicitly conditioned on every action [17]. We then define the EFE of a single time step as[2]:

$$\begin{aligned}G(s_k, o_k) &= \mathbf{KL}[q(s_k) \,||\, p(s_k, o_k)] \\ &\approx -\ln p(o_k) + \mathbf{KL}[q(s_k) \,||\, q(s_k|o_k)] \\ &\approx -r(o_k) + \mathbf{KL}[q(s_k) \,||\, q(s_k|o_k)]\end{aligned} \tag{8}$$

The negative log-likelihood (or surprise) of an observation $-\ln p(o_t)$, is replaced by the reward $-r(o_k)$ [17,7,11]. As AIF agents act to minimize their surprise, by replacing the surprise with the negative reward, we encode the agent

[2] The full derivation can be found in Appendix D

with the prior that its goal is to maximize reward. This formulation needs the EFE computation for all of the possible states and observations up to some time N, making it computationally intractable. Tractability has been achieved through bootstrapping [17,11] and combining Monte-Carlo tree search and amortized inference [5]. Here we learn a bootstrapped estimate of the EFE. The value network is used to get an estimate of the EFE for all of the possible actions. It is modelled as a feed-forward neural network $G_\rho(s_k, o_k) = f_\rho(s_k)$.

To train the value network, we use another bootstrapped EFE estimate which uses the EFE of the current time step and a $\beta \in (0, 1]$ discounted value-net estimate of the EFE under $q(a_{k+1}|s_{k+1})$ for the next time step:

$$\hat{G}(s_k, o_k) = -r(o_k) + \mathbf{KL}[q(s_k) \mid\mid q(s_k|o_k)] + \beta E_{q(a_{k+1}|s_{k+1})}[G_\rho(s_{k+1}, o_{k+1})] \tag{9}$$

Using gradient descent, we can optimize the parameters of the value network by computing the MSE between $G_\rho(s_k, o_k)$ and $\hat{G}(s_k, o_k)$:

$$L_{f_\rho,k} = MSE(G_\rho(s_k, o_k), \hat{G}(s_k, o_k)) \tag{10}$$

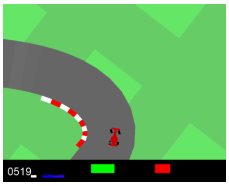
In summary, with our implementation, the VFE loss function becomes:

$$\begin{aligned} -\mathcal{F}_k =& BCE(\hat{o}_{k-N:k}, o_{k-N:k}) - \frac{1}{2}\sum(1 + \ln s_{\Sigma,k} - s^2_{\mu,k} - s_{\Sigma,k}) \\ &+ MSE(s_{\mu,k}, q_\phi(s_{k-1}, a_{k-1})) \\ &+ \mathbf{KL}[q_\xi(a_k|s_k) \mid\mid \sigma(-\gamma G_\rho(s_k, o_k))] \end{aligned} \tag{11}$$

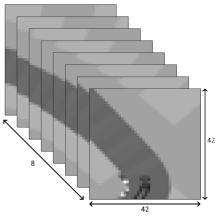
3 Experimental Setup

We evaluated the algorithm on OpenAI's CarRacing-v0 environment [13] (Fig. 2a). It is considered a Partial Observable Markov Decision Process (POMDP) problem as there is no access to the state of the agent/environment. The input is the top-view image (96×96 RGB) of part of the racing track centred on the car. The goal of this 2D game is to maximize the obtained reward by driving as fast and precise as possible. A reward of +1000/N is received for every track tile that is visited, where N is the total number of tiles (placed on the road), and a reward of -0.1 is received for every frame that passes. Solving the game entails that an agent scores an average of more than 900 points over 100 consecutive episodes. By definition, an episode is terminated after the agent visited all track tiles, the agent strayed out of bounds of the environment, or when 1000 time steps have elapsed. Every episode, a new race track is randomly generated.

The agent/car has three continuous control variables, namely steering $[-1, 1]$ (left and right), accelerating $[0, 1]$, and braking $[0, 1]$. The action space was discretized into 11 actions, similarly to [31,25,30], described in Table 2c.



(a) CarRacing-v0



(b) Preprocessed input

action	values
do nothing	[0, 0, 0]
steer sharp left	[-1, 0, 0]
steer left	[-0.5, 0, 0]
steer sharp right	[1 ,0, 0]
steer right	[0.5, 0, 0]
accelerate 100%	[0, 1, 0]
accelerate 50%	[0, 0.5, 0]
accelerate 25%	[0, 0.25, 0]
brake 100%	[0, 0, 1]
brake 50%	[0, 0, 0.5]
brake 25%	[0, 0, 0.25]

(c) Discrete Actions

4 Results

We compared our dAIF implementation with other state-of-the-art algorithms. First, Fig. 3 shows the average reward evolution while training for our dAIF architecture, Deep-Q learning (DQN) [19] (our implementation) and a random agent. Second, Table 1 shows the average reward over 100 consecutive episodes for the top methods in the literature. The average reward performance test and reward per episode for DQN and dAIF are provided in Appendix E.

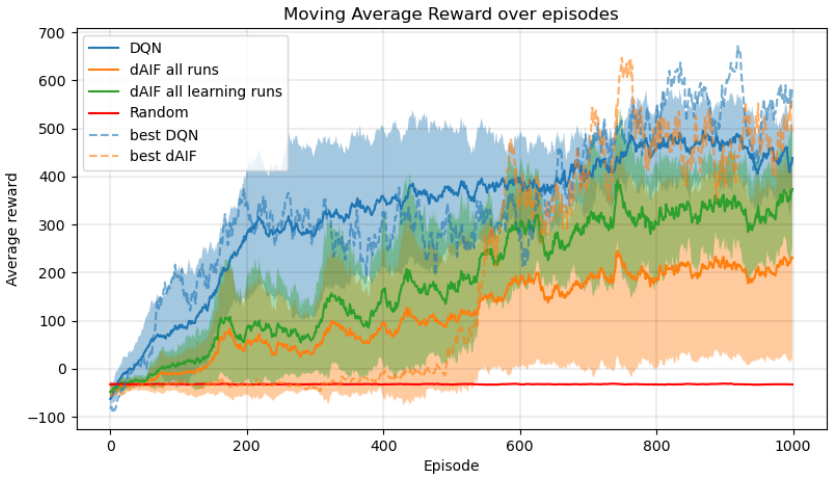


Fig. 3: Moving average reward (MAR) comparison for OpenAI's CarRacing-v0. $MAR_e = 0.1CR_e + 0.9MAR_{e-1}$, where CR_e is the cumulative reward of the current episode. In green (solid line), the mean of the dAIF runs that were able to learn a policy, and in orange (dashed line), the best of all training runs.

For the dAIF and the DQN implementations, observations are first preprocessed by removing the bottom part of the image. This part contains information about the accumulated rewards of the current episode, the car's true speed, four ABS sensors, steering wheel position, and gyroscope. Afterwards, the image is grey-scaled and reshaped to a size of 42×42. The input is defined as a stack of k to $k - N$ observations to provide temporal information by allowing the encoding of velocity and steering. We use experience replay [15] with a batch size of 250 and memory capacity of 300000 and 100000 transitions for DQN and dAIF respectively, and make use of target networks [19] (copy of the policy network for DQN and value network for dAIF) with a freeze period of 50 time steps.

The dAIF agent makes use of a pre-trained VAE which was frozen during training. Following a similar procedure as [3], the VAE was pre-trained on observations collected by having a human play the environment for 10000 time steps.

Table 1: Average rewards for CarRacing-v0

Method	Average Reward
DQN (our implementation)	515 ± 162
dAIF (our implementation)	494 ± 241
A3C (Continuous) [18]	591 ± 45
A3C (Discrete) [12]	652 ± 10
Weight Agnostic Neural Networks [9]	893 ± 74
GA [22]	903 ± 72
World models [10]	906 ± 21

5 Discussion

The dAIF implementation described in this paper has shown to reach performance on par with Deep Q-learning. However, there are some remarks. First, it showed a slower learning curve as described in previous works [11], due to the need to learn the world model. Second, we identified some runs where the system was not able to learn—See Fig. 3 orange solid line. These runs drag down the average performance. Finally, it has failed to reach state-of-the-art performance when comparing to other world model approaches—See Table 1. Here we discuss the limitations of the current implementation and alternative architectures to overcome the challenge of learning the preferences in dAIF approaches.

Observation and transition model. Our implementation does not fully exploit temporal representation learning, such as other models that use recurrent neural networks (RNN). Figure 4 describes two architectures to learn the encoding and the transitioning of the environment. Figure 4 left shows the autoencoding and transition model implemented in this work. This architecture is similar

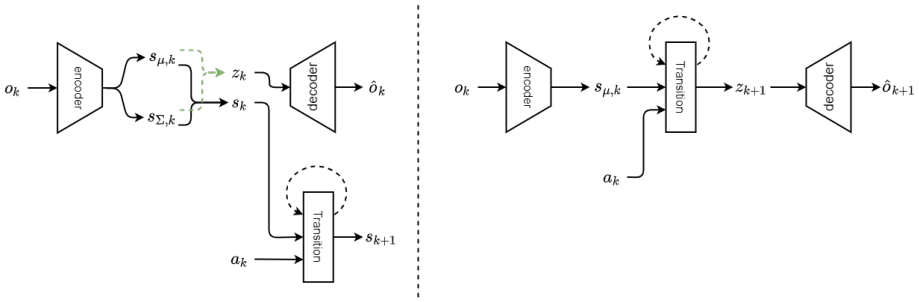


Fig. 4: Transition network outside of the observation network (left) and the transition network in between the encoder and decoder (right).

to the successful world models [10], but while we used a simple feed-forward network, they modelled the state-transition with a mixed density network RNN. Interestingly, dAIF also permits the alternative architecture described in Fig. 4 right (also proposed in [20]), in which the network learns to predict future observations. Here the transition network is part of the autoencoding. By incorporating the transition network in the structure of the observation network, we avoid the need for the dual objectives: perceptual reconstruction and dynamics learning. Preliminary testing did not show any improvements. Future work could involve more extensive testing to uncover possible performance improvements.

Dependency of input space The performance of dAIF has shown a strong dependency on the learning of the observation model. Different image pre-processing methods would lead to improvements of more than 50% in the agent performance, as shown in other DQN implementations in the literature. Testing showed that without a pre-trained observation network, the model was unable to learn consistently and rarely showed performance that would suggest an indication of task comprehension. By using a pre-trained observation network learning occurred in 6 out of the 10 runs. To produce a proper action-centric representation, likelihood, transition and control should be learnt concurrently. However, parameters uncertainty may be tackled as the models are being learnt. The current implementation uses static values for the networks learning rates, future testing could investigate different variable learning rates for each network or decaying dropout temperature.

Bootstrapping of the policy and the value. Estimating both the policy and the value from state encoding has shown end-to-end issues when we do not pre-train the observation model. In particular, 1-step ahead action formulation in conjunction with bootstrapping might not capture a proper structure of the world, which is needed to complete the task, even if we use several consecutive input images to compute the state. N-step ahead observation optimization EFE formulations, as proposed in [5,28,20], may aid learning. Particularly, when sub-

stituting the negative log surprise by the rewards, the agent might loose the exploratory AIF characteristic, thus focusing only on goal-oriented behaviour. Furthermore, and very relevant, the reward implementation in CarRacing-v0 might be not the best way to provide dAIF with rewards for proper preference learning.

References

1. Openai's carracing-v0 leaderboard. https://github.com/openai/gym/wiki/Leaderboard#carracing-v0
2. Çatal, O., Nauta, J., Verbelen, T., Simoens, P., Dhoedt, B.: Bayesian policy selection using active inference. arXiv preprint arXiv:1904.08149 (2019)
3. Çatal, O., Wauthier, S., De Boom, C., Verbelen, T., Dhoedt, B.: Learning generative state space models for active inference. Frontiers in Computational Neuroscience **14**, 103 (2020)
4. Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., Friston, K.: Active inference on discrete state-spaces: a synthesis. Journal of Mathematical Psychology **99**, 102447 (2020)
5. Fountas, Z., Sajid, N., Mediano, P.A., Friston, K.: Deep active inference agents using monte-carlo methods. arXiv preprint arXiv:2006.04176 (2020)
6. Friston, K.: A theory of cortical responses. Philosophical transactions of the Royal Society B: Biological sciences **360**(1456), 815–836 (2005)
7. Friston, K., Samothrakis, S., Montague, R.: Active inference and agency: optimal control without cost functions. Biological cybernetics **106**(8), 523–541 (2012)
8. Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J.: Action and behavior: a free-energy formulation. Biological cybernetics **102**(3), 227–260 (2010)
9. Gaier, A., Ha, D.: Weight agnostic neural networks. arXiv preprint arXiv:1906.04358 (2019)
10. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
11. van der Himst, O., Lanillos, P.: Deep active inference for partially observable mdps. In: International Workshop on Active Inference. pp. 61–71. Springer (2020)
12. Khan, M., Elibol., O.: Car racing using reinforcement learning (2018), https://web.stanford.edu/class/cs221/2017/restricted/p-final/elibol/final.pdf
13. Klimov, O.: Carracing-v0. https://gym.openai.com/envs/CarRacing-v0/
14. Lanillos, P., van Gerven, M.: Neuroscience-inspired perception-action in robotics: applying active inference for state estimation, control and self-perception. arXiv preprint arXiv:2105.04261 (2021)
15. Lin, L.: Reinforcement learning for robots using neural networks (1992)
16. Meo, C., Lanillos, P.: Multimodal vae active inference controller. arXiv preprint arXiv:2103.04412 (2021)
17. Millidge, B.: Deep active inference as variational policy gradients. Journal of Mathematical Psychology **96**, 102348 (2020)
18. Min J. Jang, S., Lee, C.: Reinforcement car racing with a3c (2017), https://www.scribd.com/document/358019044/Reinforcement-Car-Racing-with-A3C
19. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)
20. Noel, A.D., van Hoof, C., Millidge, B.: Online reinforcement learning with sparse rewards through an active inference capsule. arXiv preprint arXiv:2106.02390 (2021)

21. Parr, T., Friston, K.J.: Generalised free energy and active inference. Biological cybernetics **113**(5), 495–513 (2019)
22. Risi, S., Stanley, K.O.: Deep neuroevolution of recurrent and discrete world models. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 456–462 (2019)
23. Sancaktar, C., van Gerven, M.A., Lanillos, P.: End-to-end pixel-based deep active inference for body perception and action. In: 2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). pp. 1–8. IEEE (2020)
24. Schwartenbeck, P., Passecker, J., Hauser, T.U., FitzGerald, T.H., Kronbichler, M., Friston, K.J.: Computational mechanisms of curiosity and goal-directed exploration. Elife **8**, e41703 (2019)
25. Slik, J.: Deep reinforcement learning for end-to-end autonomous driving (2019)
26. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
27. Tschantz, A., Baltieri, M., Seth, A.K., Buckley, C.L.: Scaling active inference. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
28. Tschantz, A., Millidge, B., Seth, A.K., Buckley, C.L.: Reinforcement learning through active inference. arXiv preprint arXiv:2002.12636 (2020)
29. Ueltzhöffer, K.: Deep active inference. Biological cybernetics **112**(6), 547–573 (2018)
30. van der Wal, D., Intelligentie, B.O.K., Shang, W.: Advantage actor-critic methods for carracing (2018)
31. Zhang, Y.: Deep reinforcement learning with mixed convolutional network. arXiv preprint arXiv:2010.00717 (2020)

A Model Parameters

Table 2: General parameters

Parameter	**Value**	**Description**
$N_{screens}$	8	Size of the observation stack
N_{colour}	1	Colour channels of the input image
N_{height}	42	Height in pixels of the input image
N_{width}	42	Width in pixels of the input image
$N_{actions}$	11	Number of actions the agent can select from
$N_{episodes}$	1000	Number of episodes the model is trained for
$N_{length\_episode}$	1000	The maximum amount of time steps in an episode
Freeze period	50	The amount of time steps the target network is frozen before copying the parameters of the policy/value network
Batch size	250	Number of items in a mini-batch

Table 3: DQN parameters

Parameter	**Value**	**Description**
Policy network		Convolutional neural network which estimates Q-values given a state see Appendix B.
Target network		Copy of the Policy network which is updated after each freeze period see Appendix B.
N_{hidden}	512	Number of hidden units in the policy and target network
λ	1e-5	Learning rate
γ	0.99	Discount factor
ϵ	$0.15 \rightarrow 0.05$	Probability of selecting a random action. (Starts as 0.15, decreases linearly per episode with 0.00015 until a minimum of 0.05)
Memory capacity	300000	Number of transitions the replay memory can store

Table 4: dAIF parameters

Parameter	**Value**	**Description**
Observation network		VAE; see Appendix C.
Transition network		Feed-forward neural network of shape: $(2N_{latent} + 1)$ x N_{hidden} x $N_{actions}$
Policy network		Feed-forward neural network of shape: $2N_{latent}$ x N_{hidden} x $N_{actions}$; with a softmax function on the output
Value network		Feed-forward neural network of shape: $2N_{latent}$ x N_{hidden} x $N_{actions}$
Target network		Copy of the Value network which is updated after each freeze period
N_{hidden}	512	Number of hidden units in the transition, policy, and value network.
N_{latent}	128	Size of the latent state
$\lambda_{transition}$	1e-3	Learning rate of the transition network
λ_{policy}	1e-4	Learning rate of the policy network
λ_{value}	1e-5	Learning rate of the value network
λ_{VAE}	5e-6	Learning rate of the VAE
γ	12	Precision parameter
β	0.99	Discount factor
α	18000	$\frac{1}{\alpha}$ is multiplied with the VAE loss to scale its size to that of the other term in the VFE
Memory capacity	100000	Number of transitions the replay memory can store

B DQN: Policy network

Table 5: Layers DQN policy network

Type	out channels	kernel	stride	input	output
conv	64	4	2	(1, 8, 42, 42)	(1, 64, 20, 20)
batchnorm					
maxpool	-	2	2	(1, 64, 20, 20)	(1, 64, 10, 10)
relu					
conv	128	4	2	(1, 64, 10, 10)	(1, 128, 4, 4)
batchnorm					
maxpool	-	2	2	(1, 128, 4, 4)	(1, 128, 2, 2)
relu					
conv	256	2	2	(1, 128, 2, 2)	(1, 256, 1, 1)
relu					
dense	-	-	-	256	512
dense	-	-	-	512	11

C VAE

Table 6: VAE layers

Type	out channels	kernel	stride	input	output	
conv	32	4	2	(1, 8, 42, 42)	(1, 32, 20, 20)	Encoder
batchnorm						
relu						
conv	32	4	2	(1, 32, 20, 20)	(1, 64, 9, 9)	
batchnorm						
relu						
conv	128	5	2	(1, 64, 9, 9)	(1, 128, 3, 3)	
batchnorm						
relu						
conv	256	3	2	(1, 128, 3, 3)	(1, 256, 1, 1)	
relu						
dense	-	-	-	256	128	
dense μ	-	-	-	128	128	
dense logΣ	-	-	-	128	128	
dense	-	-	-	128	128	Decoder
dense	-	-	-	128	256	
deconv	128	3	2	(1, 256, 1, 1)	(1, 128, 3, 3)	
batchnorm						
relu						
deconv	64	5	2	(1, 128, 3, 3)	(1, 64, 9, 9)	
batchnorm						
relu						
deconv	32	4	2	(1, 64, 9, 9)	(1, 32, 20, 20)	
batchnorm						
relu						
deconv	8	4	2	(1, 32, 20, 20)	(1, 8, 42, 42)	
batchnorm						
relu						
sigmoid						

D Derivations

Derivation for the EFE for a single time step:

$$
\begin{aligned}
G(s_k, o_k) &= \mathbf{KL}[q(s_k) \mid\mid p(s_k, o_k)] \\
&= \int q(s_k) \ln \frac{q(s_k)}{p(s_k, o_k)} \\
&= \int q(s_k) \ln q(s_k) - \ln p(s_k, o_k) \\
&= \int q(s_k) \ln q(s_k) - \ln p(s_k|o_k) - \ln p(o_k) \\
&\approx \int q(s_k) \ln q(s_k) - \ln q(s_k|o_k) - \ln p(o_k) \\
&\approx -\ln p(o_k) + \int q(s_k) \ln q(s_k) - \ln q(s_k|o_k) \\
&\approx -\ln p(o_k) + \int q(s_k) \ln \frac{q(s_k)}{q(s_k|o_k)} \\
&\approx -\ln p(o_k) + \mathbf{KL}[q(s_k) \mid\mid q(s_k|o_k)] \\
&\approx -r(o_k) + \mathbf{KL}[q(s_k) \mid\mid q(s_k|o_k)]
\end{aligned}
$$

E Average Reward over 100 episodes

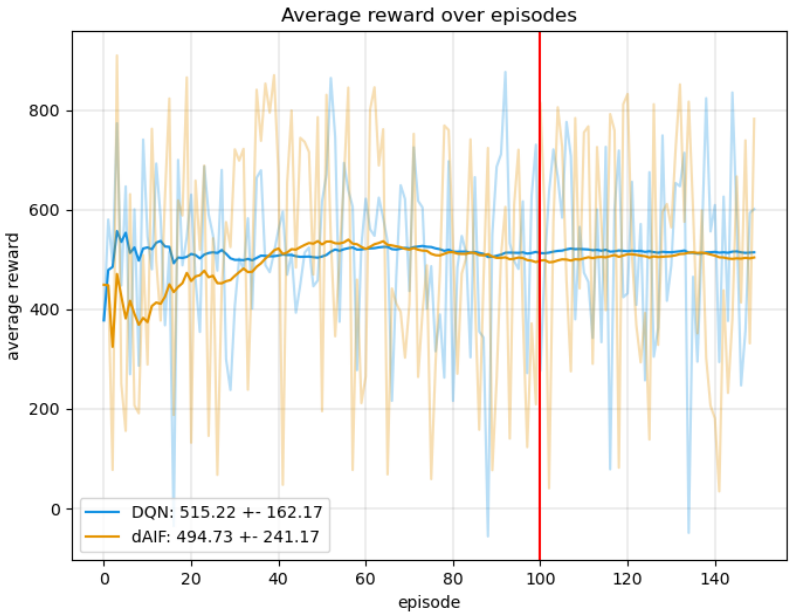


Fig. 5: Average reward test over 100 episodes for DQN and dAIF. The bright lines show the mean over episodes. The transparent lines show the reward that was obtained in a particular episode.

Robot Localization and Navigation through Predictive Processing using LiDAR

D. Burghardt[1] and P. Lanillos[2]

[1] Radboud University, Houtlaan 4, 6525 XZ Nijmegen, NL
[2] Donders Institute for Brain, Cognition and Behaviour, Department of Artificial Intelligence, Radboud University, Nijmegen, NL

Abstract. Knowing the position of the robot in the world is crucial for navigation. Nowadays, Bayesian filters, such as Kalman and particle-based, are standard approaches in mobile robotics. Recently, end-to-end learning has allowed for scaling-up to high-dimensional inputs and improved generalization. However, there are still limitations to providing reliable laser navigation. Here we show a proof-of-concept of the predictive processing-inspired approach to perception applied for localization and navigation using laser sensors, without the need for odometry. We learn the generative model of the laser through self-supervised learning and perform both online state-estimation and navigation through stochastic gradient descent on the variational free-energy bound. We evaluated the algorithm on a mobile robot (TIAGo Base) with a laser sensor (SICK) in Gazebo. Results showed improved state-estimation performance when comparing to a state-of-the-art particle filter in the absence of odometry. Furthermore, conversely to standard Bayesian estimation approaches our method also enables the robot to navigate when providing the desired goal by inferring the actions that minimize the prediction error.

Keywords: Predictive Processing · Robot localization · Robot navigation · Laser sensor · LiDAR.

1 Introduction

Localization algorithms are part of our daily life and core for robotics. Recursive Bayesian estimation composes the current state-of-art in the field and has been essential for the development of localization, mapping, navigation and searching applications [20, 11]. Bayesian filters [5], e.g., Kalman and particle filters, are able to estimate the state of a system from noisy sensor observations formalized as a hidden Markov model. These approaches are useful also in the case of non-linear modeled systems and out-of-sequence measurements [3]. The particle filter (PF) is an approximate Bayesian method that tractably computes the posterior distribution of the state of any system given the observations. The state distribution is represented by individual particles, which are evaluated and weighted recursively. Particles with higher probability get bigger weights and are re-sampled into more particles in its neighborhood, whereas particles with smaller weights get fewer new samples close to them [14].

In recent years, novel approaches based on deep neural networks, have been proposed to improve localization using high-dimensional inputs. Regression solutions, for instance, compute the absolute position of the system from only visual information [9]. However, these methods have lower accuracy than previous approaches that exploit prior information, such as geometry [19]. In particular, LiDAR-based Navigation with representation learning (e.g., using autoencoders) and reinforcement learning has shown downgraded performance in navigation tasks [8]. Alternatively to LiDAR-based approaches, however, the work of [21] has demonstrated a successful application of deep active inference in robot navigation using camera images.

We describe how the predictive processing (PP) approach to perception [4, 12] can aid in localization and simple navigation tasks [13]. In this work navigation is performed having the robot move between two points in an unobstructed environment, which can be further built upon to tackle more complex environments (e.g. mazes). Under PP, the agent, following the Free Energy Principle (FEP) [6], tries to minimize the error in the predicted observations by either performing corrective actions to match the expected internal state or by updating this internal state based on what it has experienced through the senses. In this work, we present a proof-of-concept based on the Pixel-Active Inference model [18], proposed for humanoid body perception and action, to perform laser-based localization and navigation without the need for odometry. This has been successfully applied to robot manipulator control to improve adaptation [15]. Our approach combines the power of deep networks regression with variational Bayesian filtering to provide a better reliable state estimation than PFs in our proof-of-concept environment—See Fig. 1.

2 Methods

2.1 Robot

The TIAGo Base mobile robot uses the SICK TiM571 laser sensor, which has a $0.0\text{m} - 25\text{m}$ range and a $270°$ aperture angle. In all experiments we limited the robot's movement to 2 degrees of freedom, i.e., moving forward, backward and sideways.

2.2 Localization

We define the true state (position) of the robot at instant k as $\mathbf{x}_k = (x, y) \in \mathbb{R}^2$ and the position belief of the robot as $\tilde{\mathbf{x}}_k$. We further define the observation $\mathbf{o}_k$ as the laser measurements. Estimation is solved by computing the posterior distribution $p(\mathbf{x}|\mathbf{o})$ by optimizing the Variational Free Energy (VFE). The algorithm is sketched in Fig. 2a. Under the mean-field and Laplace approximation this is equivalent to minimizing the error between the sensory input $\mathbf{o}_k$ and the predicted sensory input $\hat{\mathbf{o}}_k$. While regression approaches (Fig. 2b) compute the

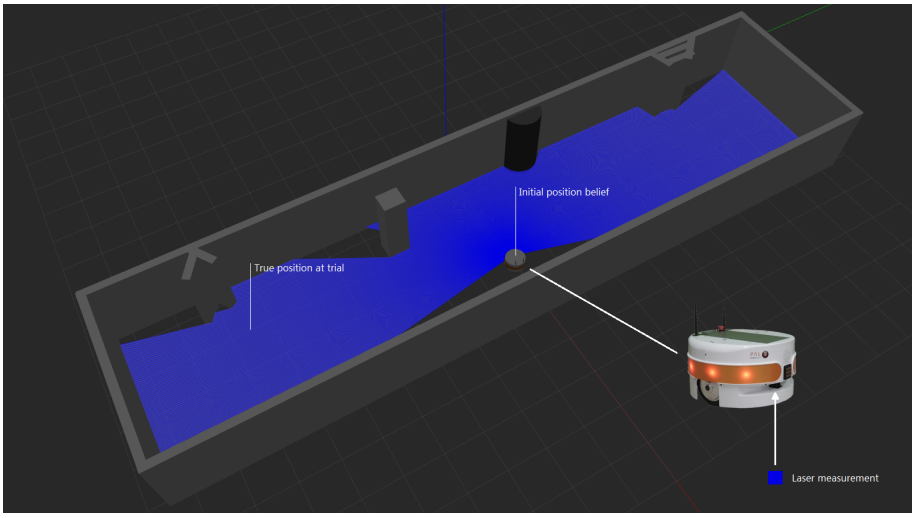


Fig. 1: Proof-of-concept environment designed for the experiments in Gazebo. In the localization experiment the robot true position is randomly set and the initial position belief is initialized to the center of the map. Laser range finder measurements are displayed as the blue shading.

pose directly from the visual input, stochastic neural filtering continuously refines the state through an error signal. We perform state estimation through perceptual inference, minimizing the VFE as follows:

$$\tilde{\mathbf{x}} = \underset{\tilde{\mathbf{x}}}{\operatorname{argmin}} F(\tilde{\mathbf{x}}, \mathbf{o}) \rightarrow \tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{x}}_k + \alpha \partial_{\tilde{\mathbf{x}}} g(\tilde{\mathbf{x}}_k) \Sigma_{\mathbf{o}}^{-1} (\mathbf{o}_k - g(\tilde{\mathbf{x}}_k)) \tag{1}$$

Where α is the step size and $\partial_{\tilde{\mathbf{x}}}$ denotes the derivative with respect to $\tilde{\mathbf{x}}$. This is computed iteratively using gradient descent on the prediction error—sensor measurement $\mathbf{o}_k$ minus the predicted sensory input $g(\tilde{\mathbf{x}}_k)$—weighted by the variance $\Sigma_{\mathbf{o}}$. Both the predicted observations and the partial derivative of the error are computed by means of a deep neural network forward pass and its Jacobian [18], respectively.

2.3 Predicting the observations

We compute the sensor likelihood $p(\mathbf{o}_k|\tilde{\mathbf{x}}_k)$ using a transposed convolutional neural network (Fig. 2), which augments the dimensionality of the input from $\mathbf{x}$ to the laser-sensor input size (e.g., $2 \rightarrow 622$). The input is firstly fed into two fully connected layers, and each transposed convolution layer is followed by a regular convolution layer, based on the work of [18]. At every layer, we used the ReLU activation function.

The network was trained on 13000 normalized random samples collected in the Gazebo simulation. Each sample consists of the true position of the robot

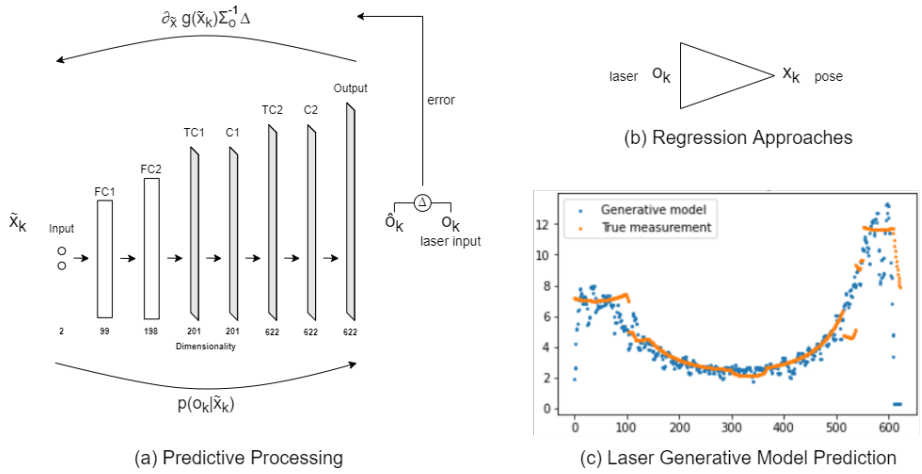


Fig. 2: (a) The Predictive Processing algorithm's architecture. (b) Graphical representation of regression approaches to robot state estimation. (c) Generative model prediction vs. true laser measurement on a test sample.

and the laser-sensor measurements at that location. The training was performed with 20 batches of 500 samples, using the L1 loss and Adam optimizer [10].

2.4 Navigation

Analogously, our algorithm infers the action in the same way that state estimation is computed, namely performing *active inference*. Actions also minimize the VFE in the predicted observations.

$$\mathbf{a} = \underset{\mathbf{a}}{\operatorname{argmin}}\, F(\tilde{\mathbf{x}}, \mathbf{o}(\mathbf{a})) \tag{2}$$

We define the goal as the preference or the intention of the agent $\tilde{\mathbf{x}}_{goal}$ to arrive to a sensory state o_{goal} [17]. Estimation and control are computed as follows[3]:

$$\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{x}}_k + \alpha \left[\partial_{\tilde{\mathbf{x}}} g(\tilde{\mathbf{x}}_k)\Sigma_{\mathbf{o}}^{-1}(\mathbf{o}_k - g(\tilde{\mathbf{x}}_k)) + \partial_{\tilde{\mathbf{x}}} g(\tilde{\mathbf{x}}_k)\Sigma_{\mathbf{x}}^{-1}\beta(\mathbf{o}_k - g(\tilde{\mathbf{x}}_{goal}))\right] \tag{3}$$

$$\mathbf{a}_{k+1} = \mathbf{a}_k + \gamma \partial_a \tilde{\mathbf{x}} \partial_{\tilde{\mathbf{x}}} g(\tilde{\mathbf{x}}_k)\Sigma_{\mathbf{o}}^{-1}(\mathbf{o}_k - g(\tilde{\mathbf{x}}_k)) \tag{4}$$

where β weights the goal attractor and γ is the action step size. Note that each term computes the weighted prediction error mapped to the latent space.

The estimated state $\tilde{\mathbf{x}}$, now biased by the desired goal, generates a new predicted observation at every new iteration that is transformed into an action $\mathbf{a}$, which minimizes the VFE. Thus, performing a movement in the direction of the goal. The pseudo-code described in Alg. 1 illustrates the process. The algorithm converges when the observation fits the predicted laser sensor measurements.

[3] This update equation assumes that the Hessian of the goal dynamics is -1 as proposed in [18].

Algorithm 1: FEP localization and navigation algorithm

$\tilde{\mathbf{x}} \leftarrow$ initial belief;
$\mathbf{o}_{goal} \leftarrow g(\tilde{\mathbf{x}}_{goal})$; `// Generate goal`
while *true* **do**
  $\mathbf{o}_k \leftarrow$ Normalize(laser input);
  $\hat{\mathbf{o}}_k \leftarrow g(\tilde{\mathbf{x}})$; `// Predicted observation`
  $\tilde{\mathbf{x}} \leftarrow$ Eq. 3;
  $\mathbf{a} \leftarrow$ Eq. 4;
  PerformAction(a);
end

3 Results

We evaluated our laser-based Active Inference algorithm against a particle filter [14] in the Gazebo simulator, using a commercial mobile robot with a laser rangefinder sensor (TIAGo Base, pmb-2) [2], interfaced with Robot Operating System (ROS) [1]. All experiments were conducted in a designed corridor-like map described in Fig. 1. Localization and navigation results are summarized in Fig. 3.

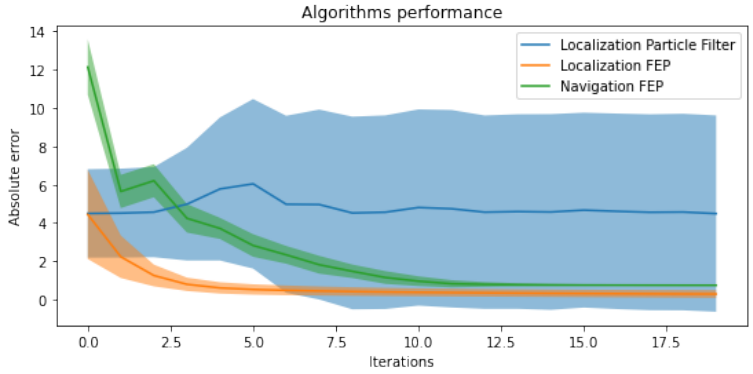


Fig. 3: Localization and navigation evaluation. Mean and standard deviation of the positioning absolute error of our model (FEP) compared against the PF algorithm. Furthermore, our model is able to perform navigation using the same Bayesian filtering framework. The green line shows the mean absolute error to the goal.

3.1 Localization and estimation

Firstly, we evaluated the localization accuracy when initializing the robot to a random position in the space when a map was given. Thus, testing absolute

positing with laser measurements in static situations. For the particle filter, the particle initial probabilities were randomly spread in the environment and reset at the beginning of every trial. In our algorithm, we initialized the initial belief in the center of the environment. We computed the ground truth positional error in 100 trials for 50 iterations. Our algorithm converged much more consistently to the true state over all trials, whereas the PF struggled to deliver consistent results, as shown in Fig. 3 by the rather large standard deviation in the blue shaded area. It is important to highlight that the PF is tuned for using the robot's odometry. However, for the sake of fair comparison solely laser-sensor values were used as observations.

Secondly, we evaluated the localization performance when traversing the environment from one side to the other, by performing small teleports (to override odometry) to simulate robot movement while keeping the rotation angle constant. Results showed a more stable over time state estimation by our algorithm when compared to the PF, which seemed to suffer from the absence of odometry information.

3.2 Navigation

For the assessment of the navigation algorithm's performance, we ran an experiment consisting of 50 trials in which the robot had to navigate from a starting point to a goal position. The initial belief state was set to the robot's initial true position, to evaluate the performance of navigation without the effects of localization in the first iterations. In every trial, both the initial position and the goal state of the robot were chosen randomly, with the constraint that they should be at least 12 meters (in Gazebo coordinates) apart from each other. The task was considered complete when the robot got in a range of 0.8m from the target. The results are plotted in green in Fig. 3.

We observed that the robot initially quickly approximates the goal, with a big drop in the distance to the goal in the first couple of iterations. As it gets closer to the goal, the "velocity" of the robot (in the experiment described by the step sizes) decreases. This is a result of the diminishing gradient in every step of the algorithm, due to the stochastic gradient descent. Additionally, we computed the average number of iterations that it took the algorithm to get in the desired 0.8m range of the target. Over the 50 trials of similar travel distance ($\sim 11m$ to $\sim 13.5m$), the average number of iterations was 12.5. This number is naturally closely related to the optimal step size found.

4 Conclusions

This work shows a proof-of-concept on how predictive processing, i.e. active inference agents, can perform laser-based localization and navigation tasks. The results obtained in the localization experiment, where we compared our approach against a state-of-the-art alternative (particle filter), show the potential of predictive stochastic neural filtering in robot localization, and estimation in

general [7, 16]. Furthermore, the navigation experiment showcased how to compute actions as a dual filtering process. Nevertheless, at its current state, the proposed algorithm suffers from a few deficiencies, most of which are related to the learning of the generative model of the world. Besides currently requiring a large dataset for training, the model is prone to mistake very similar objects in the environment, given that the estimation of the new state is independent from the previous. Additionally, because it is a supervised method trained before that the robot can do any navigation, it is unable to cope with changing environments. All in all, the environment used in our experiments is rather simplistic compared to demonstrations of current sota algorithms. Therefore, we foresee further development and experimentation in terms of integration of odometry information, the introduction of extra degrees of freedom and connection to the robot's non-linear dynamics.

References

1. Ros documentation (Jun 2020), http://wiki.ros.org/
2. Tiago base (Oct 2020), http://wiki.ros.org/Robots/TIAGo-base
3. Besada-Portas, E., Lopez-Orozco, J.A., Lanillos, P., De la Cruz, J.M.: Localization of non-linearly modeled autonomous mobile robots using out-of-sequence measurements. Sensors **12**(3), 2487–2518 (2012)
4. Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences **36**(3), 181–204 (2013)
5. Fox, V., Hightower, J., Liao, L., Schulz, D., Borriello, G.: Bayesian filtering for location estimation. IEEE Pervasive Computing **2**(3), 24–33 (2003). https://doi.org/10.1109/MPRV.2003.1228524
6. Friston, K.: The free-energy principle: a unified brain theory? Nature reviews neuroscience **11**(2), 127–138 (2010)
7. Friston, K.J., Trujillo-Barreto, N., Daunizeau, J.: Dem: a variational treatment of dynamic systems. Neuroimage **41**(3), 849–885 (2008)
8. Gebauer, C., Bennewitz, M.: The pitfall of more powerful autoencoders in lidar-based navigation. arXiv preprint arXiv:2102.02127 (2021)
9. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
11. Lanillos, P.: Minimum time search of moving targets in uncertain environments. Ph.D. thesis, PhD thesis (2013)
12. Lanillos, P., Cheng, G.: Adaptive robot body learning and estimation through predictive coding. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4083–4090. IEEE (2018)
13. Lanillos, P., van Gerven, M.: Neuroscience-inspired perception-action in robotics: applying active inference for state estimation, control and self-perception. arXiv preprint arXiv:2105.04261 (2021)
14. Liu, B., Cheng, S., Shi, Y.: Particle filter optimization: A brief introduction pp. 95–104 (2016)
15. Meo, C., Lanillos, P.: Multimodal vae active inference controller. arXiv preprint arXiv:2103.04412 (2021)

16. Millidge, B., Tschantz, A., Seth, A., Buckley, C.: Neural kalman filtering. arXiv preprint arXiv:2102.10021 (2021)
17. Oliver, G., Lanillos, P., Cheng, G.: An empirical study of active inference on a humanoid robot. IEEE Transactions on Cognitive and Developmental Systems (2021)
18. Sancaktar, C., van Gerven, M.A.J., Lanillos, P.: End-to-end pixel-based deep active inference for body perception and action. 2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (Oct 2020). https://doi.org/10.1109/icdl-epirob48136.2020.9278105, http://dx.doi.org/10.1109/ICDL-EpiRob48136.2020.9278105
19. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixe, L.: Understanding the limitations of cnn-based absolute camera pose regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3302–3312 (2019)
20. Thrun, S.: Simultaneous localization and mapping. In: Robotics and cognitive approaches to spatial mapping, pp. 13–41. Springer (2007)
21. Çatal, O., Wauthier, S., Verbelen, T., Boom, C.D., Dhoedt, B.: Deep active inference for autonomous robot navigation (2020)

Sensorimotor Visual Perception on Embodied System Using Free Energy Principle

Kanako Esaki[0000-0002-3269-9130], Tadayuki Matsumura, Kiyoto Ito[0000-0002-2243-5756], and Hiroyuki Mizuno[0000-0002-1213-9021]

Research & Development Group, Hitachi, Ltd.,
1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo, 185-8601, Japan
kanako.esaki.oa@hitachi.com

Abstract. We propose an embodied system that is based on the free energy principle (FEP) for sensorimotor visual perception (SMVP). Although the FEP mathematically describes the rule that living things obey, limitation by embodiment is required to model SMVP. The proposed system is configured by a body, which partially observes the environment, and memory, which retains classified knowledge about the environment as a generative model, and executes active and perceptual inferences. Evaluation using the MNIST dataset showed that the proposed system recognizes characters by active and perceptual inferences, and the intentionality corresponding to human confirmation bias is reproduced on the system.

Keywords: Free energy, Embodiment, Sensorimotor contingency.

1 Introduction

The human visual field seems to cover a wide area of the surrounding environment, but the range with high enough resolution to identify details is limited to only the central visual field of about 5 degrees [1]. Since the human visual field has this spatial limitation, gazing-position movement is required to see the environment. Sensorimotor contingency (SMC) theory [2,3] changes the interpretation of "seeing" [4–9] by including this movement in it. SMC theory explains that "seeing" is knowing about things to do rather than making an internal representation [10]. This means that human sensorimotor visual perception (SMVP) includes moving the gazing position using the inference of the environmental state, not the sensory input itself, to understand the environment [11]. The spatial limitation of the human visual field within this context is not a "limitation" but a "trigger" for an action that moves the gazing position[12].

Many machine-learning methods have been proposed that incorporate human characteristics. Spatial limitations, such as that described above, are treated as constraints called partial observation in the context of reinforcement learning [13–15]. These studies evaluated the performance degradation of classification and regression through partial observation. Various methods similar to SMVP have also been proposed. Auto regressive models [16–18] predict the entire image by repeating the action of obtaining a

partial image. Algorithms for reinforcement learning [19,20] generate exploring actions covering a wide range of the action space. These models and algorithms generate the actions on the basis of sensory inputs such as partial images and observations. The active vision algorithms infer hidden states from partial observations and finally understand the whole scene, but the hidden state space is constructed in terms of reconfigurabilty [21], or actions are selected on a basis of image features instead of hidden states [22–25]. None of the above studies have used the inference of environmental states to generate action, which is the essence of SMVP.

The purpose of this study is to achieve SMVP using the free energy principle (FEP) [26–30]. The FEP mathematically describes the rule that living things obey. The free energy in the FEP measures the difference between the probability distribution of environmental states that act on a biological system and an approximate posterior distribution of environmental states encoded by the configuration of that system. The biological system minimizes the free energy by changing its configuration to affect the way it samples the environment or by changing the approximate posterior distribution it encodes. These two changes correspond to "active inference" and "perceptual inference" of environmental states, respectively.

Although the FEP mathematically describes active and perceptual inferences on a biological system, limitation by embodiment of such a system to trigger action is required to model SMVP. Embodiment [31–34] provides an interaction between the biological system and environment, resulting in partial sensory inputs and actions. Their causal relationship is condensed to be stored in the embodied biological system.

We propose an embodied system that is based on the FEP to achieve SMVP. The proposed system is configured by a body, which partially observes the environment, and memory, which retains condensed knowledge about the environment. Evaluation using the MNIST dataset [35] showed that the proposed system triggers an action that moves a gazing position and repeatedly executes active and perceptual inferences by following the FEP. Moreover, the intentionality is reproduced on the proposed system, producing an equivalent of human confirmation bias. We discuss how important this bias is for taking the action in an unknown environment.

2 Sensorimotor Visual Perception

The problem settings shown in Fig. 1 are designed to list the components necessary for SMVP. Let us consider a situation in which a target object, e.g., the number 5, exists in the environment. This situation is called an environmental state x_t. The vision sensor takes on the role of the human eye and has a spatial limitation of the visual field. This limitation leads to a change in direction of the vision sensor to understand the environment. The vision sensor obtains an image of a specific region of the environment each time its direction is determined. A representative position of the region is defined as an attention position, which equals the gazing position of the human eye. The image of the region is defined as an attention image. The attention image at each time ($T = t-2, \mathrm{t}-1, \mathrm{t}$) is obtained by the previous actions $a_{t-3}, a_{t-2}, a_{t-1}$ that move the attention position. A composition image of the attention images obtained from $T = 0$ to $T = t$ is

used as a sensory input (hereafter called sensory input image s_t). In our problem setting described above, SMVP is defined to infer the environment states by repeating the actions to obtain sensory input images.

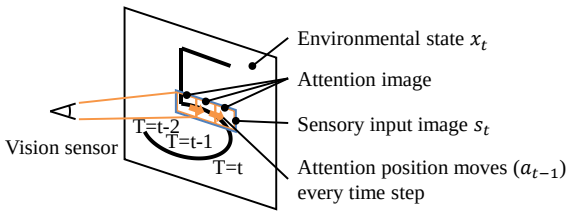


Fig. 1. Overview of problem setting to consider sensorimotor visual perception

3 Free Energy Principle

To apply the FEP to our problem setting, we make certain assumptions. The s_t is determined by only the a_{t-1} and the current x_t. The x_t does not change ($x_{t-1} = x_t = x_{t+1}$).

Under the above assumptions, the variational and expected free energy of the FEP, which are necessary components to describe the perceptual and active inferences, are expressed as follows. The variational free energy $F(\phi_{x_t}, a_{t-1})$ is expressed as

$$\begin{aligned} F(\phi_{x_t}, a_{t-1}) &= E_{q(x_t|\phi_{x_t})}\left[\ln q(x_t|\phi_{x_t}) - \ln p_{a_{t-1}}(x_t, s_t)\right] \\ &= D_{KL}\left[q(x_t|\phi_{x_t})||p_{a_{t-1}}(x_t|s_t)\right] - \ln p_{a_{t-1}}(s_t)\,, \end{aligned} \tag{1}$$

where $q(x_t|\phi_{x_t})$ is the approximate posterior distribution of x_t, ϕ_{x_t} is the sufficient statistics of $q(x_t)$, and $p_{a_{t-1}}(x_t, s_t)$ is a generative model that stores the causal relationship of x_t and s_t under a_{t-1}. Since a_{t-1} specifies the prior distribution of s_t, it is treated as a parameter of the generative model. Perceptual inference is aimed at minimizing $F(\phi_{x_t}, a_{t-1})$ by changing $q(x_t|\phi_{x_t})$. Since the second term of Eq. 1 is composed of a_{t-1} and s_t, which are fixed at the current time t, the purpose is achieved by changing ϕ_{x_t} to become $q(x_t|\phi_{x_t}) \sim p_{a_{t-1}}(x_t|s_t)$. The expected free energy $G(\phi_{x_{t+1}}, a_t)$, on the other hand, is expressed as

$$
\begin{aligned}
G(\phi_{x_{t+1}}, a_t) &= E_{q(s_{t+1}|x_{t+1},\phi_{x_{t+1}})} F(\phi_{x_{t+1}}, a_t) \\
&= E_{q(s_{t+1}|x_{t+1},\phi_{x_{t+1}})} E_{q(x_{t+1}|\phi_{x_{t+1}})} \left[\ln q(x_{t+1}|\phi_{x_{t+1}}) - \ln p_{a_t}(x_{t+1}, s_{t+1})\right] \\
&= E_{q(s_{t+1}|x_{t+1},\phi_{x_{t+1}})} \Big[E_{q(x_{t+1}|\phi_{x_{t+1}})} \left[\ln q(x_{t+1}|\phi_{x_{t+1}}) - \ln p(x_{t+1})\right] \\
&- E_{q(x_{t+1}|\phi_{x_{t+1}})} \ln p_{a_t}(s_{t+1}|x_{t+1})\Big] \\
&= E_{q(s_{t+1}|x_{t+1},\phi_{x_{t+1}})} D_{KL}\left[q(x_{t+1}|\phi_{x_{t+1}})||p(x_{t+1})\right] \\
&+ E_{q(s_{t+1}|x_{t+1},\phi_{x_{t+1}})} \Big[-E_{q(x_{t+1}|\phi_{x_{t+1}})} \ln p_{a_t}(s_{t+1}|x_{t+1})\Big],
\end{aligned}
\tag{2}
$$

where $p_{a_t}(x_{t+1}, s_{t+1})$ is factorized into $p_{a_t}(s_{t+1}|x_{t+1})p(x_{t+1})$ since x_{t+1} and a_t are independent. Since s_{t+1} varies with a_t but x_{t+1} does not in our problem setting, $p_{a_t}(x_{t+1}, s_{t+1})$ is factorized into the terms with and without s_{t+1}, unlike the general active inference literature[29]. Active inference is aimed at minimizing $G(\phi_{x_{t+1}}, a_t)$ by changing action a_t. Since a_t is included only in the second term (hereafter called uncertainty) and the first term is fixed, the purpose is achieved by minimizing the uncertainty.

4 Embodied System for Sensorimotor Visual Perception

Limitation by an embodiment is key to achieving SMVP based on the FEP. The proposed embodied system is configured by a body and memory. The body has an ocular motor system for controlling the attention position. In our problem setting shown in Fig. 1, the vision sensor, which has a spatial limitation of the visual field, is the body. The body thus can only observe a partial area of the environment. The memory, on the other hand, is not a photographic memory, where x_t is observed uniformly as if it were photographed. Rather, it is a generative model that contains classified prior knowledge about the causal relationship of an s_t and x_t to be stored in a limited capacity. By limiting body and memory abilities and operating in accordance with the FEP, the proposed embodied system repeatedly executes perceptual and active inferences, as shown in Fig. 2. In perceptual inference, an s_t is generated and input to the generative model $p_{a_{t-1}}(x_t, s_t)$ to calculate an approximate posterior. In active inference, the uncertainty is calculated using the approximate posterior and used to select an attention position.

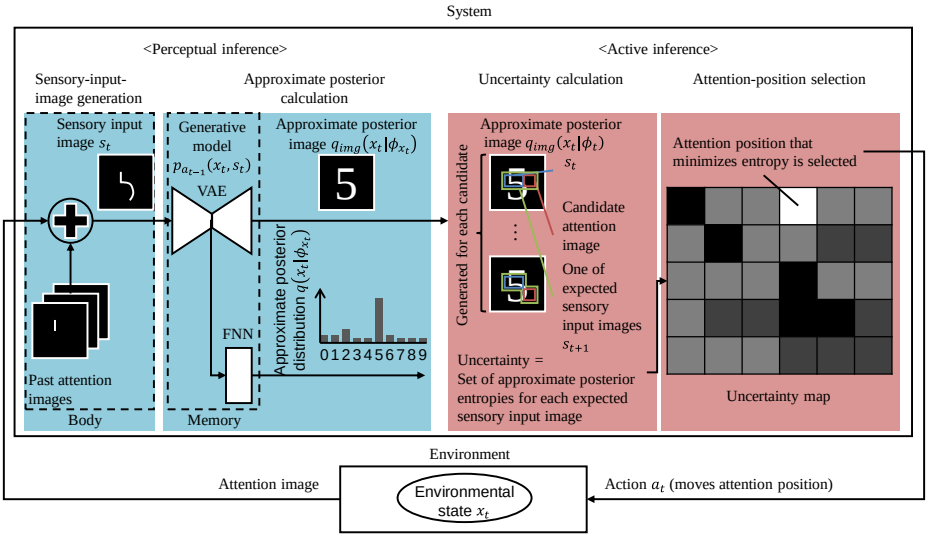


Fig. 2. Processing flow of sensorimotor visual perception with proposed embodied system

4.1 Perceptual Inference

The generation of s_t involves a current attention image obtained from the environment and the past attention images obtained from the initial time to the current time t. The attention images are composed while maintaining their relative attention positions.

The generative model $p_{a_{t-1}}(x_t, s_t)$ is implemented with a combination of a variational autoencoder (VAE) and a fully connected neural network (FNN), which is inspired by the auxiliary loss in GoogleNet [36]. The VAE contains prior knowledge about the environment, and the FNN classifies it. The combination of these is pretrained, which is equivalent to changing ϕ_{x_t} to become $q(x_t|\phi_{x_t}) \sim p_{a_{t-1}}(x_t|s_t)$ in Eq. 1. An approximate posterior distribution $q(x_t|\phi_{x_t})$ and approximate posterior image $q_{img}(x_t|\phi_{x_t})$, which is a conversion of $q(x_t|\phi_{x_t})$ into an image format, are calculated by inputting the s_t to the combination of the VAE and FNN.

4.2 Active Inference

Calculation of uncertainty uses $q_{img}(x_t|\phi_{x_t})$. Uncertainty is the expected information amount of $p_{a_t}(s_{t+1}|x_{t+1})$, which is the probability distribution of the set of expected sensory input images s_{t+1} conditioned by x_{t+1} with action a_t as a parameter. Conditioning by $x_{t+1}(= x_t)$ is interpreted as extracting s_{t+1} from $q_{img}(x_t|\phi_{x_t})$. Each one of s_{t+1} is composed of the current s_t and candidate attention image surrounding it. All the s_{t+1} are extracted from one single $q_{img}(x_t|\phi_{x_t})$. Parameterizing a_t is interpreted as assuming the candidate attention images. Since the action space is continuous in the temporal direction under limited body, candidate images should be limited to the region

surrounding s_t. Under these interpretations, the information amount of $p_{a_t}(s_{t+1}|x_{t+1})$ is calculated as that of s_{t+1}. The information amount of s_{t+1} is the entropy of the approximate posterior distribution calculated by inputting each one of s_{t+1} to $p_{a_{t-1}}(x_t, s_t)$. Since s_{t+1} is deterministically calculated from one $q_{img}(x_t|\phi_{x_t})$, the expected-value calculation is not required. Uncertainty is thus the set of the entropies for each one of s_{t+1}.

In selecting the attention position, an uncertainty map is generated where the entropy for each one of s_{t+1} corresponds to each attention position. An attention position that minimizes the entropy is selected in accordance with the uncertainty map. The a_t moves the attention position to the selected one.

5 Evaluation and Discussion

We evaluated the proposed embodied system for SMVP in a character-recognition task with the MNIST dataset. The key components of the enbodiment, body and memory, were implemented as follows: the body function was achieved by setting the size of the attention image to be sufficiently small compared to the MNIST characters, and the memory (the generative model) was implemented using convolutional VAE [37,38] that was robust against displacement. The size of both the s_t and approximate posterior image $q_{img}(x_t|\phi_{x_t})$ was 28 × 28 pixels, while that of the attention image was 6 × 6 pixels. Training of the generative model involved 60,000 images of the MNIST dataset. The stochastic variables x_t of the approximate posterior distribution $q(x_t|\phi_{x_t})$ were labels 0–9 of the MNIST dataset.

Perceptual inference was executed during the attention repetitions. Figure 3 shows the transition of $q(x_t|\phi_{x_t})$ and $q_{img}(x_t|\phi_{x_t})$ from the 1st to 20th attentions when target characters are “0” and “3”. Symbols A, B, C, and D in Fig. 3 (a) are described later. Each graph plots the $q(x_t|\phi_{x_t})$ and each image is the corresponding $q_{img}(x_t|\phi_{x_t})$. The probability of the characters different from the target characters changed to the maximum. This situation is analogous to a human temporarily labeling the environmental state so that they could identify the environment on the basis of the sensory input they had collected and their past experience. For character “0”, the probability of the target character was maximum after the 13th attention, and “0” appeared in the $q_{img}(x_t|\phi_{x_t})$. For character “3”, however, the probability of “2” was maximum even at the 20th attention, and the $q_{img}(x_t|\phi_{x_t})$ contained a part of “3” not the entire “3”.

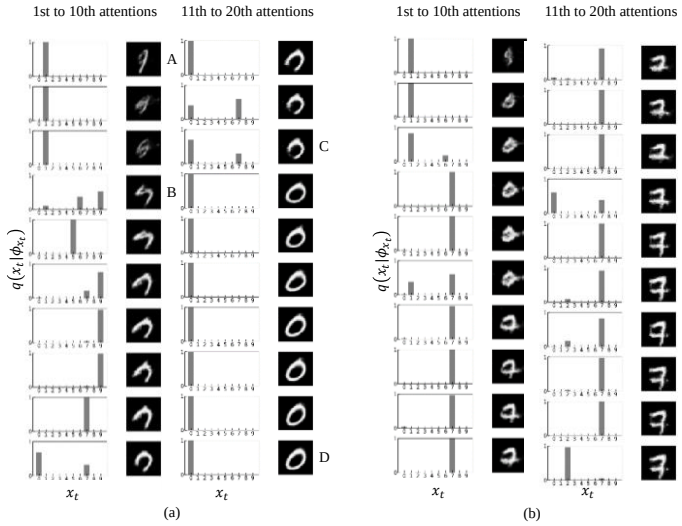


Fig. 3. Transition of approximate posterior distributions and images along with attention repetitions. Subfigures (a) and (b) correspond to making attention at characters "0" and "3", respectively.

Active inference was executed during attention repetitions. Figure 4 shows the uncertainty map at the 1st, 4th, 13th, and 20th attentions (corresponding to A, B, C, and D in Fig. 3(a)) when the target character is "0". The vertical (x) and horizontal (y) axes of the maps indicate the representative position of the candidate attention image. The uncertainty of the 1st and 20th attentions were biased toward the minimum because the change in uncertainty is similar for any next action. In the early stage of attention repetition, any next action will be useful for inference since little environmental information has been collected. In the final stage, the probability of a particular number has been high, and the entropy of $q(x_{t+1}|\phi_{x_{t+1}})$ is low with any next action. The uncertainty of the 4th and 13th attentions were dispersed because any of the next actions has the potential to specify a number.

The proposed embodied system executed SMVP in which the attention position moves on the basis of the inference of x_t. Figure 5 shows the transition of attention images, s_t, and $q_{img}(x_t|\phi_{x_t})$ from the 1st to 20th attentions. Each subset of three rows corresponds to the target character from "0" to "9". The 1st, 2nd, and 3rd rows show the attention images, s_t, and $q_{img}(x_t|\phi_{x_t})$, respectively. Before the 9th attention of target character "2", a form like "7" was inferred and the attention position moved up and down, but then a form like "2" was inferred and the attention position moved to the right. Before the 15th attention of target character "5", a form like "6" was inferred and the attention position moved around the bottom of x_t, but then a form like "5" was inferred and the attention position moved to the upper right.

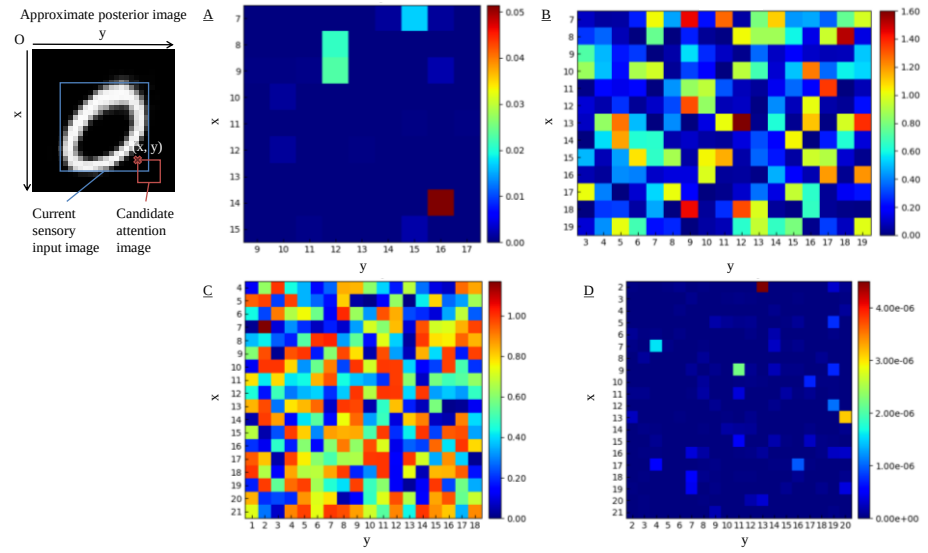


Fig. 4. Uncertainty map when making attention at "0". Subfigures A, B, C, and D correspond to 1st, 4th, 13th, and 20th attentions, respectively.

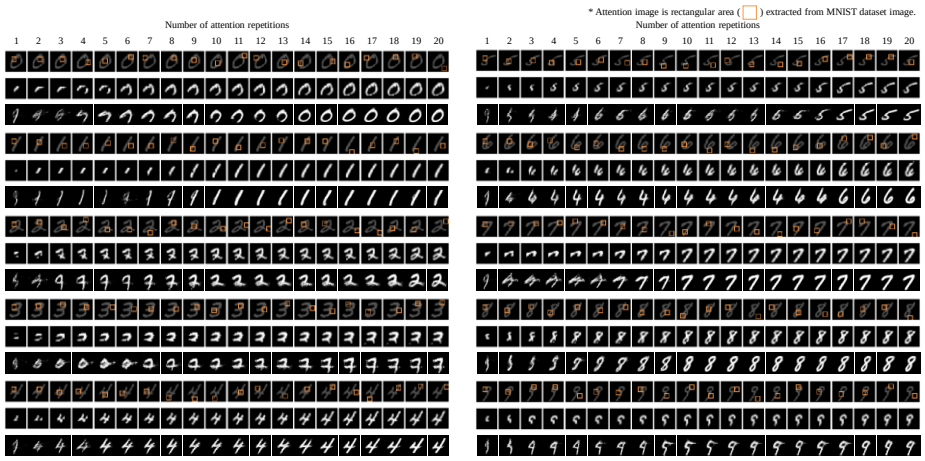


Fig. 5. Transition of attention images* (first row), sensory input images (second row), and approximate posterior images (third row) along with number of attention repetitions. Each subset of three rows corresponds to making attentions from "0" to "9".

To analyze the case of character "3", the initial attention position was changed. Figure 6 shows the transition of the attention image, s_t, $q(x_t|\phi_{x_t})$, and $q_{img}(x_t|\phi_{x_t})$ from the 1st to 20th attentions when the initial attention position was different from those in Figs 3 and 5 for character "3". Different from Fig. 3 (b), the probability of "3" reached maximum, and the $q_{img}(x_t|\phi_{x_t})$ contained the entire "3" at the 20th attention. This result suggests that the proposed system has an intentionality similar to human

confirmation bias that depends on what is obtained from the environment and prior knowledge about it. Although human confirmation bias has a negative impact on decision-making in various fields [39], it has the advantage of adaptability to unknown environments. In most practical cases of perception problems, incomplete models of the environment are provided. In these cases, humans take the next action on the basis of the confirmation bias. Taking the next exploring action enables the environmental information to be obtained. We believe that our results will help solve the difficult problem of triggering action in an unknown environment.

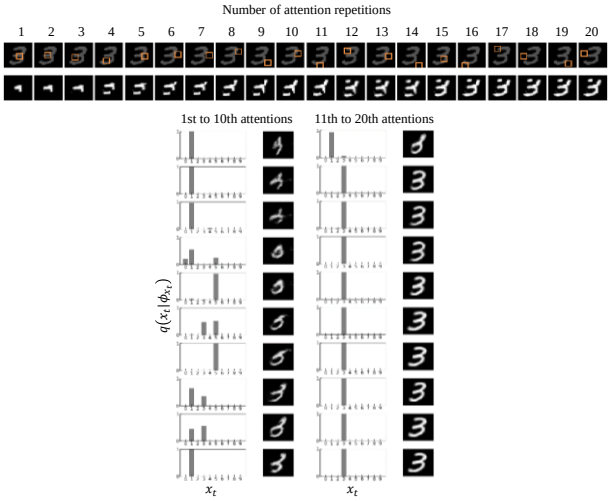


Fig. 6. Transition of attention images, sensory input images, and approximate posterior. Changes in initial attention position when making attention at "3" provides different transition

6 Conclusion

We proposed an embodied system that is based on the free energy principle (FEP) for sensorimotor visual perception (SMVP). The proposed embodied system is configured by a body and memory. By limiting body and memory abilities and operating in accordance with the FEP, the proposed system triggers an action that moves an attention position and repeatedly executes perceptual and active inferences. During the evaluation involving a character-recognition task using the MNIST dataset, as the attention was repeated, the uncertainty of the characters decreased. The probability of the correct character finally became the highest among the characters. It was thus confirmed that the proposed system greatly contributes to achieving SMVP. Moreover, changing the initial attention position provides a different final inference, suggesting that the proposed system has a confirmation bias similar to humans. We believe that these results will help solve the difficult problem of triggering action in an unknown environment.

Acknowledgements. The authors thank Dr. Qinghua Sun from Hitachi Ltd. for his constructive comments and suggestions for improving this paper.

Appendix

The generative model, described in this paper, is a combination of a variational autoencoder (VAE) and a fully connected neural network (FNN). The architecture is shown in Fig. 7. The encoder consists of four 2D convolutional layers and each layer is followed by a batch normalization and a rectified linear unit. The bottleneck consists of two linear transformation layers for calculating the average and the variance with reparameterizing function. The decoder consists of four 2D transposed convolutional layers and each layer is followed by a batch normalization and a rectified linear unit (sigmoid unit for the last layer). The classifier consists of a linear transformation layer followed by a rectified linear unit and a linear transformation layer followed by a softmax unit. The model was trained using Adam optimizer (learning rate: 0.001) with the sum of VAE loss and FNN loss.

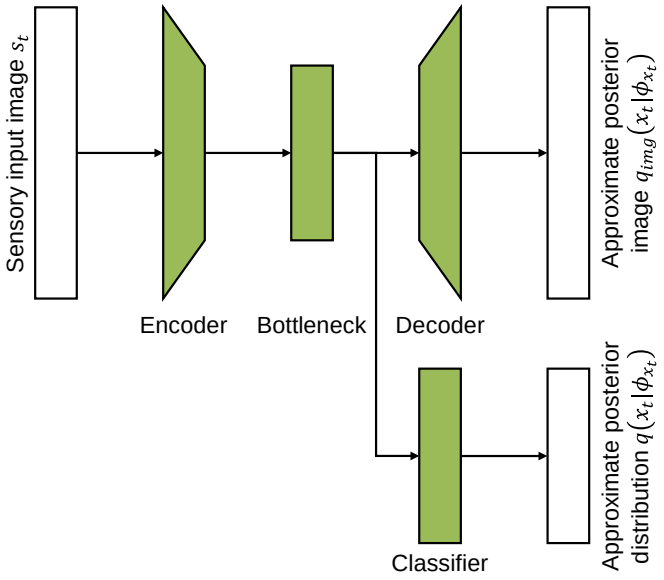


Fig. 7. Architecture of generative model

Algorithm 1 shows the pseudo code of the processing flow. The $p_{a_{t-1}}(s_t, x_t)$ is pre-trained using training data of (s_t, x_t). All the training data of s_t are pre-processed so that the center of gravity of an image is shifted to the center position. During the operation of the proposed embodied system, the process from the 2nd line to the 12th line is repeated. First, an attention image s'_t is obtained from the vision sensor. The past sensory input images are composed with the obtained s'_t while maintaining each relative attention position. The center of gravity of the composed image is calculated, and the composed image is shifted so that the center of gravity is located at the center position of the image. The shifted composed image is an s_t. Then, $q(x_t|\phi_{x_t})$ is calculated by inputting s_t to $p_{a_{t-1}}(s_t, x_t)$. After that, the sub-function starting from the 14th line

is called to generate expected sensory input images s_{t+1}. In the sub-function, an $q_{img}(x_t|\phi_{x_t})$ is calculated by inputting s_t to $p_{a_{t-1}}(s_t, x_t)$. A template image is generated by detecting a bounding rectangle area of non-zero pixels in s_t and extracting the area from s_t. Template matching is carried out in the $q_{img}(x_t|\phi_{x_t})$, and the representative position of the current s_t, u_{cur}, is obtained. To calculate the next candidate attention positions u_{next}, a candidate region of $u_{next}$is set. The candidate region is a region obtained by adding a fixed margin pixel to a region of s_t in $q_{img}(x_t|\phi_{x_t})$. The region of s_t is defined by u_{cur} and the size of the template image. The u_{next} are calculated by sliding the window with the fixed stride pixel in the candidate region. The window is the size of s'_t. The representative positions of all the window positions during sliding are u_{next}. The s_{t+1} are generated by extracting the region of the s_t and the region of the next candidate attention images s'_{t+1} from $q_{img}(x_t|\phi_{x_t})$. The region of the s_t is defined by u_{cur} and the size of the template image, as mentioned above. The region of the s'_{t+1} are defined by u_{next} and the size of s'_{t+1}. The extracted images are clipped or applied with zero-padding to have the same size as s_t. Each approximate posterior distribution $q(x_{t+1}|\phi_{x_{t+1}})$ is calculated by inputting each image included in s_{t+1} to $p_{a_{t-1}}(s_t, x_t)$. Note that $q(x_t|\phi_{x_t})$ is calculated using the current s_t, while $q(x_{t+1}|\phi_{x_{t+1}})$ is calculated using s_{t+1}. The entropy of each $q(x_{t+1}|\phi_{x_{t+1}})$ is calculated and added to the uncertainty map M. Finally, the attention position having the minimum value in M is defined as the next attention position.

Algorithm 1 Pseudo code of processing flow

```
while system is operating do:
    obtain attention image s'_t
    compose past sensory input images with s'_t
    generate sensory input image s_t by shifting gravity point to center of
    composed image
    calculate approximate posterior distribution q(x_t|φ_{x_t}) using s_t and
    generative model p_{a_{t-1}}(s_t, x_t)
    call generate_expected_sensory_input_image
    for expected sensory input images s_{t+1}:
        calculate approximate posterior distribution q(x_{t+1}|φ_{x_{t+1}}) using
        s_{t+1}[index] and p_{a_{t-1}}(s_t, x_t)
        calculate entropy of q(x_{t+1}|φ_{x_{t+1}})
        add entropy to uncertainty map M
    end for
    set attention position using M
end while

generate_expected_sensory_input_image
    generate approximate posterior image q_img(x_t|φ_{x_t}) using s_t
    and p_{a_{t-1}}(s_t, x_t)
    generate template from s_t
```

obtain current sensory input position u_{cur} by template matching in $q_{img}(x_t|\phi_{x_t})$
calculate next candidate attention positions u_{next} using u_{cur}
generate s_{t+1} using $q_{img}(x_t|\phi_{x_t})$, u_{cur}, and u_{next}
return with s_{t+1}

References

1. Mandelbaum, J., Sloan, L. L.: Peripheral visual acuity*: with special reference to scotopic illumination. American Journal of Ophthalmology 30(5), 581–588 (1947).
2. O'Regan, J. K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences 24(5), 939–973 (2001).
3. Seth, A. K.: The cybernetic Bayesian brain: from interoceptive inference to sensorimotor contingencies. Open MIND (35), (2015).
4. Land, M. F.: Eye movements and the control of actions in everyday life. Progress in Retinal and Eye Research 25(3), 296–324 (2006).
5. Friston, K., Kiebel, S.: Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B Biological Sciences 364(1521), 1211–1221 (2009).
6. Seth, A. K., Suzuki, K., Critchley, H. D.: An interoceptive predictive coding model of conscious presence. Frontiers in Psychology 2, 395 (2012).
7. Adams, R. A., Shipp, S., Friston, K. J.: Predictions not commands: active inference in the motor system. Brain Structure and Function 218(3), 611–643 (2013).
8. Bogacz, R.: A tutorial on the free-energy framework for modelling perception and learning. Journal of Mathematical Psychology 76(Part B), 198–211 (2017).
9. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. In: 5th International Conference on Learning Representations, Toulon (2017).
10. O'Regan, J. K.: Experience is not something we feel but something we do: a principled way of explaining sensory phenomenology, with Change Blindness and other empirical consequences, http://nivea.psycho.univ-paris5.fr/ASSChtml/Pacherie4.html, last accessed 2021/8/27.
11. Parr, T., Sajid, N., Da Costa, L., Mirza, M. B., Friston, K. J.: Generative models for active vision. Frontiers in Neurorobotics 15, 34 (2021).
12. Tang, Y., Nguyen, D., Ha, D.: Neuroevolution of self-interpretable agents. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference, pp. 414–424. Association for Computing Machinery, Cancún (2020).
13. Pineau, J., Gordon, G., Thrun, S.: Point-based value iteration: an anytime algorithm for POMDPs. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 1025–1030. Morgan Kaufmann Publishers Inc., Acapulco (2003).
14. Ji, S., Parr, R., Li, H., Liao, X., and Carin, L.: Point-based policy iteration. In: Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2, pp. 1243–1249. AAAI Press, Vancouver (2007).
15. Silver, D., Veness, J.: Monte-Carlo planning in large POMDPs. Advances in Neural Information Processing Systems 23, 2164–2172 (2010).

16. Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., Wierstra, D.: DRAW: a recurrent neural network for image generation. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 1462–1471. JMLR.org, Lille (2015).
17. Oord, A. V., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: Proceedings of the 33rd International Conference on Machine Learning, pp. 1747–1756. JMLR.org, New York (2016).
18. Salimans, T., Karpathy, A., Chen, X., Kingma, D. P.: PixelCNN++: Improving the pixelCNN with discretized logistic mixture likelihood and other modifications. In: 5th International Conference on Learning Representations, Toulon (2017).
19. Oh, J., Guo, X., Lee, H., Lewis, R., Singh, S.: Action-conditional video prediction using deep networks in Atari games. Advances in Neural Information Processing Systems 28, 2863–2871 (2015).
20. Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., Abbeel, P.: VIME: variational information maximizing exploration. Advances in Neural Information Processing Systems 29, 1117–1125 (2016).
21. van der Himst, O., Lanillos, P.: Deep active inference for partially observable MDPs. In: Verbelen, T., Lanillos, P., Buckley, C. L., De Boom, C. (eds.) International Workshop on Active Inference 2020, Communications in Computer and Information Science, vol. 1326, pp. 61–71. Springer (2020).
22. Daucé, E., Perrinet, L.: Visual search as active inference. In: Verbelen, T., Lanillos, P., Buckley, C. L., De Boom, C. (eds.) International Workshop on Active Inference 2020, Communications in Computer and Information Science, vol. 1326, pp. 165–178. Springer (2020).
23. Friston, K., Adams, R. A., Perrinet, L., Breakspear, M.: Perceptions as hypotheses: saccades as experiments. Frontiers in Psychology 3, 151 (2012).
24. Mirza, M. B., Adams, R. A., Mathys, C. D., Friston, K. J.: Scene construction, visual foraging, and active inference. Frontiers in Computational Neuroscience 10, 56 (2016).
25. Heins, R. C., Mirza, M. B., Parr, T., Friston, K., Kagan, I., Pooresmaeili, A.: Deep active inference and scene construction. Frontiers in Artificial Intelligence 3, 81 (2020).
26. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. Journal of Physiology-Paris 100(1–3), 70–87 (2006).
27. Friston, K.: The free-energy principle: a unified brain theory?. Nature Reviews Neuroscience 11, 127–138 (2010).
28. McGregor, S., Baltieri, M., Buckley, C. L.: A minimal active inference agent. arXiv preprint arXiv:1503.04187 (2015).
29. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: a process theory. Neural Computation 29(1), 1–49 (2017).
30. Buckley, C. L., Kim, C. S., McGregor, S., Seth, A. K.: The free energy principle for action and perception: a mathematical review. Journal of Mathematical Psychology 81, 55–79 (2017).
31. Fitzpatrick, P., Metta, G., Natale, L., Rao, S., Sandini, G.: Learning about objects through action - initial steps towards artificial cognition. In: 2003 IEEE International Conference on Robotics and Automation, pp. 3140–3145. IEEE, Taipei (2003).
32. Cheng, G., Hyon, S., Morimoto, J., Ude, A., Colvin, G., Scroggin, W., Jacobsen, S. C.: CB: a humanoid research platform for exploring neuroscience. In: 2006 6th IEEE-RAS International Conference on Humanoid Robots, pp. 182–187. IEEE, Genova (2006).
33. Friston, K.: Embodied inference: or "I think therefore I am, if I am what I think". In: Tschacher , W., Bergomi , C. (eds.) The implications of embodiment: Cognition and communication, pp. 89–125. Imprint Academic (2011).

34. Gallagher, S., Allen, M.: Active inference, enactivism and the hermeneutics of social cognition. Synthese 195, 2627–2648 (2018).
35. THE MNIST DATABASE of handwritten digits, http://yann.lecun.com/exdb/mnist/, last accessed 2021/8/27.
36. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. IEEE, Boston (2015).
37. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, Banff (2014).
38. Rezende, D. J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1278–1286. JMLR.org, Beijing (2014).
39. Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., Sharot, T.: Confirmation bias in the utilization of others' opinion strength. Nature Neuroscience 23, 130–137 (2020).