Active Inference - A Possible Planning Mechanism in the Human Cognitive Apparatus?

MSc. Cognitive and Decision Sciences

Candidate Number: BPWF6

A Thesis Presented for the Degree of

Master of Science

Supervisor: Prof. David Lagnado

University College London

31.08.2023

Permission to share thesis with future MSc CoDeS students granted

MSc Cognitive and Decision Sciences
Project Assessment Form

Candidate Number: BPWF6

Name of Supervisor(s): Prof. David Lagnado

Do you consent to make your anonymised essay available as an example or a learning resource for future students? Yes

Please use the table below in conjunction with the **PALS Guide for Class Descriptors for PGT Project Marking** which has been sent to you.

Please note that you are not required to give numerical grades in the table below, just tick the relevant box

Grade	80%+	70-79%	60-69%	50-59%	40-49%	Below 40%
Classification	**High Distinction**	**Distinction**	**Merit**	**Pass**	**Fail**	**Fail**
General						
Literature Review						
Method						
Results						
Discussion						

Feedback comments:

FINAL GRADE	

Abstract

Humans are constantly faced with uncertainty. A cognitive apparatus that is confronted with a complex and dynamically changing environment thus benefits from being equipped with mechanisms that can deal with uncertainty efficiently. This thesis investigates whether active inference, a process theory of the free-energy framework, is a plausible candidate for such mechanisms. In essence, the free-energy principle posits that all (living) matter strives to minimize uncertainty in order to survive. We hypothesized that an active inference agent, which we specifically designed for the present thesis, fits human behavioural data better than classic reinforcement learning algorithms that deal with uncertainty, namely Thompson sampling and the Upper Confidence Bound algorithm as well as a randomly performing agent. The task design was a restless bandit task taken from Gershman (2018). We used Bayesian inference with unknown mean and unknown variance as the learning rule for all algorithms, while the above-mentioned algorithms served as choice rules. Based on maximum likelihood estimation, we indeed found that active inference models human behaviour better than the comparison algorithms. These results, while only being a first step in that direction, shed light on the computational processes that might be involved in planning and agency.

1 Introduction

Natural dynamics are complex and hard to predict. This means that humans are constantly confronted with uncertainty, making efficient navigation of stochasticity an important factor of evolutionary fitness. This is a key insight that has led to theories about the Bayesian brain receiving much scholarly attention in recent years (see e.g., Clark, 2013; Hohwy, 2020; Seth, 2021). Bayesian theories provide the benefit that by making uncertainty a vital parameter of inference and choice, they deal with it in a principled and quantified way. The latter seems necessary as uncertainty reveals information about reinforcement that can be received from the environment (Speekenbrink, 2022).[1]

Most interactions of humans with the environment are not one-shot games, but sequential processes where feedback can be integrated into subsequent interaction cycles. This means that an agent at any given time faces a dilemma between seeking more information to reduce uncertainty or exploiting what has already been learned to maximise some sort of goal or reward - the agent is thus confronted with the so-called exploration-exploitation dilemma. In this thesis, we will investigate how the free-energy principle and its process theory, active inference, deal with this. In essence, the free-energy principle

[1] This point will be elaborated in detail below.

posits that from single cells to complex organisms, all living beings strive to minimize entropy. It has received much scholarly attention in recent years, ranging from philosophy, neuroscience or machine learning to even robotics (Clark, 2013; Kirchhoff et al., 2018; Millidge, 2021). For example, researchers have recently investigated the usefulness of active inference algorithms in letting robot landers make more autonomous decisions in remote regions such as space missions (Wakayama and Ahmed, 2023).

In the endeavour to empirically test whether this theory applies to humans, we created an active inference reinforcement learning agent and let it solve two-armed restless bandit tasks. Specifically, we believe that active inference agents are computationally and behaviourally plausible and hence, besides showcasing good task performance, are a good model of human behaviour. While much has been said about the biological plausibility (Friston, 2009, 2010) and validity of the active inference approach (Friston et al., 2015; Parr et al., 2022; Tschantz et al., 2020), a comparison of an active inference agent and human behaviour in the classic sequential decision making paradigm of bandit tasks has thus far been lacking. Furthermore, we are not aware of any active inference algorithms designed to solve a bandit with an assumed normal distribution. This project aims to rectify these two omissions.

This research is important as it can provide insights into plausible computational mechanisms used by the brain. The proposed argument of this thesis is one of plausibility - if it is accepted that active inference is computationally plausible and if we establish that in an experiment which reflects real-world learning and decision making scenarios active inference algorithms correlate highly (and more highly than alternative reinforcement learning algorithms) with human performance, then it is plausible that the brain actually does something similar to active inference. Of course, even in the event of such a positive finding this is not conclusive evidence, but should rather be seen as fundamental research that points to fruitful avenues for future research.

In the following, in section 2, we will first motivate Bayesian theories of the brain and how they are connected to the free-energy principle and its process theory, active inference. Section 3 then introduces the exploration-exploitation dilemma which arises when an agent is confronted with planning an action in an uncertain environment and how expected free-energy tries to deal with this problem. Lastly, in section 3.3, we introduce the environment in which such planning processes can be tested - the multi-armed bandit task. Section 4 then presents the specific implementation of the two-armed bandit task used in this thesis as well as the Bayesian inference learning rule common to all algorithms we compare in this work. Section 4.3 then presents the choice rules that are compared, namely active inference, Upper Confidence Bound (UCB), Thompson sampling and a randomly choosing agent. Section 4.4

and 4.5 subsequently introduce the instructions the participants in the original study received as well as the main analysis used in this work, namely maximum likelihood estimation. The results section is separated into descriptive and inferential statistics whereby the first section discusses checks of validity of our algorithms and the inferential statistics section presents the result of the main analysis, which is that active inference, as hypothesized, does model human data better than the comparison algorithms. In section 6, we discuss that this might be due to its variational approach to risk assessment or due to it encoding an information gain term explicitly rather than implicitly.[2]

2 Bayes, Free-Energy Principle and Active Inference

2.1 The Bayesian Background

The derivation of the free-energy principle starts with the idea that perception and action are fundamentally inference problems. This idea dates back to Helmholtz and Kant, who discovered that there are barriers between the self and the outside world (Helmholtz, 1860; Kant, 1999). Kant developed this point in his transcendental idealism and differentiated between *noumenon*, an object that exists independently of the senses and *phenonmenon*, a reception of the senses of the noumenal world (Kant, 1999). Importantly, he states that the noumenal world is not sensible and can thus not directly be known to us - our mental machinery can merely access an image of the world in itself (Kant, 1999).

Helmholtz developed this point by hypothesizing about the mental machinery involved. According to his view, the brain has no direct access to the outside world and can only infer what is going on in the environment (which includes bodily states) via electrical input received from its sensory organs (Helmholtz, 1860). If one takes into account the fact that these electrical firing patterns do not have a one-to-one mapping to stimuli in the environment (one stimulus can lead to different firing patterns and the same firing pattern can follow from different stimuli), the fundamental task of the brain is probabilistic inference - coming up with a robust mapping of firing patterns and environmental states. This process is assumed to be top-down driven rather than bottom-up. This does not mean that there is no bottom-up communication, but that the model of the world is communicated top-down. That is to say sensory information is used to confirm/disconfirm hypotheses about the data generating process, rather than creating the phenomenon in an inductive way with a causal chain from the outside world to the retina to the relevant cortical areas, leading eventually to the phenomenon, or perception.

[2] The code for the environment, agents as well as data analysis can be found here: https://github.com/marvin-math/active_inference_multiarmed_bandit.

Bayesian theories of the brain specify how this inference process can go about. While not the only approach to deal with uncertainty (Colombo et al., 2021), it has been argued to be the normative standard of rationality (Clark, 2013; Hohwy, 2013, 2020). Indeed, the benefits of the Bayesian approach seem intuitively appealing: information can be sequentially integrated into a system of beliefs, meaning that each subsequent interaction with the environment is more informed than the previous one. Furthermore, Bayesianism is based on subjective probabilites, or - in our context - generative models, which reflect the Helmholtzian assumption that perception is not a bottom-up, but a top-down process and is thus inherently subjective (Hohwy, 2013). This subjectivity does not mean that the probability updating process is not connected in any way to the outside world - it is controlled by and in a constant feedback relation with the stimuli provided by the outside world (Seth, 2021). However, the initial assumptions are not directly provided by the outside world, but rather by the agent.

It is hence not surprising that there is a wealth of evidence suggesting that mental phenomena rest on Bayesian computations. For example, it is likely that Bayesian processes play a role in visual perception, as phenomena like background clutter or illumination cause visual scenes to be highly ambiguous and complex. These complexities and ambiguities can be dealt with by probabilistically integrating prior object knowledge with perceived information (Kersten et al., 2004). Furthermore, there is a vast literature on the involvement of Bayesian processes in illusions of several senses (Clark, 2013; Ehrsson, 2007; Lenggenhager et al., 2007; Olivé and Berthoz, 2012; Seth, 2021). Besides visual illusions there are also so-called body-ownership illusions, in which body-ownership is assigned to locations in space outside of the actual body (Olivé and Berthoz, 2012). Presumably the most famous one in this category is the rubber hand illusion, where the participant's real hand is hidden and a rubber hand is placed in good sight of the participant. Both, the real and the rubber hand, are then synchronously stroked, resulting in people assigning sensations to the rubber hand and in some cases even reports of mislocalizing the actual hand towards the rubber hand (Botvinick and Cohen, 1998).

A parsimonious explanation for such illusions would again be the involvement of Bayesian updating rules and a prominent role of priors in perception (Seth, 2021). This is because the expectations built in the above and similar experiments seem to make one interpretation of the sensory input more likely than another. For example, it has been argued that the rubber hand illusion arises because visual information usually takes primacy and is thus trusted more in terms of localization of the feeling (Clark, 2013; Seth, 2021). In this thesis, we will take a Bayesian point of view of belief updating by assuming that the process of updating the model of the world works in a Bayesian fashion. Specificially, this means that the learning rule for the agents developed in this thesis follows Bayesian update rules. However, some Bayes-optimal computations overwhelm our computational resources, which has been a major motivation for the development of the free-energy principle.

2.2 The Free-Energy Principle

The free-energy principle asserts that in order to survive, biological agents have to avoid a natural tendency to disorder (Friston, 2009, 2010). In other words, according to the free-energy principle, the essential task of any biological system is to occupy a limited set of states within a viable range of values in order to ensure its survival. For example, humans should avoid the physiological state of having a body temperature of over 41 degrees Celsius, as this is not a state humans can survive in for very long. Another way of saying this is that this state has high entropy. Entropy is the long-term average of surprise, which is a measure of uncertainty of an outcome. Surprise can be formulated as the negative log evidence of observed data under the generative model of an agent, $-\log p(\mathcal{X})$, and in order to ensure homeostasis, agents aim to minimise this term.[3] Note that a state that is not to be preferred has just been translated into an improbable state. This translation is a centerpiece of the free-energy framework: because organisms tend to occupy states that ensure their survival and these are states of low entropy/surprise or high model evidence, we can translate a state that is beneficial just as a state that is likely to be occupied.

However, surprise (or model evidence) can be intractable, rendering the above surprise minimization problem impossible or too hard to compute. For most real-world problems, computing surprise would involve marginalizing out too many hidden states, given that biological agents operate with limited computational resources. This is why it is assumed that the agent performs variational approximations. What this essentially means is that the posterior distribution is approximated with tractable and less complex distributions. This converts an intractable integration problem into an optimization problem (Parr et al., 2022). To see how this works, we start with Bayes Theroem:

$$P(\vec{\Theta}|x) = \frac{P(x|\vec{\Theta})P(\vec{\Theta})}{P(x)} \tag{1}$$

$$P(x)P(\vec{\Theta}|x) = P(x|\vec{\Theta})P(\vec{\Theta}) \tag{2}$$

[3] The term surprise here does not denote the feeling of being surprised, but rather refers to states that are improbable. A classic example is a fish outside of water. While this state might not be accompanied by any mental state of surprise by the fish, it has high surprisal in the sense of being a state that is not viable for long for the fish and should be avoided - and is therefore improbable (Friston, 2010).

For a continuous variable we can say that:

$$P(x) = \int P(x, \vec{\Theta}) d\vec{\Theta} \tag{3}$$

With $\vec{\Theta}$ being a placeholder for parameters and latents. Here we can see the above-mentioned problem: computing the model evidence would require marginalising over all $\vec{\Theta}$, which is usually not feasible in practice. To convert this into an optimisation problem, we first multiply the term inside the integral with an arbitrary variational distribution over the latent states and parameters $Q(\vec{\Theta})$ divided by itself, which results in multiplying by one and does not change anything mathematically.

$$p(x) = \int \frac{Q(\vec{\Theta})}{Q(\vec{\Theta})} p(x, \vec{\Theta}) d\vec{\Theta} \tag{4}$$

$$p(x) = \int Q(\vec{\Theta}) \frac{p(x, \vec{\Theta})}{Q(\vec{\Theta})} d\vec{\Theta} \tag{5}$$

We then take the log of both sides.

$$\log p(x) = \log \int Q(\vec{\Theta}) \frac{p(x, \vec{\Theta})}{Q(\vec{\Theta})} d\vec{\Theta} \tag{6}$$

Now we can make use of Jensen's inequality, which states that the log of an average is always greater than or equal to the average of a log (Parr et al., 2022). We take advantage of this inequality by pushing the log inside to turn it into an expectation of the log of the joint over the variational distribution and a lower bound on the log evidence.

$$\log p(x) \geq \int Q(\vec{\Theta}) \log \frac{p(x, \vec{\Theta})}{Q(\vec{\Theta})} d\vec{\Theta} \tag{7}$$

After multiplying both sides by -1, we have an upper bound on the negative log evidence, which we call variational free-energy, F.

$$-\log p(x) \leq -\int Q(\vec{\Theta}) \log \frac{p(x, \vec{\Theta})}{Q(\vec{\Theta})} d\vec{\Theta} = F \tag{8}$$

By factorising the joint, we can then express the free-energy as follows

$$\begin{aligned}
F &= -\mathbb{E}_{Q(\vec{\Theta})}\left[\log \frac{p(x,\vec{\Theta})}{Q(\vec{\Theta})}\right] && (9)\\
&= -\mathbb{E}_{Q(\vec{\Theta})}\left[\log \frac{p(\vec{\Theta}|x)p(x)}{Q(\vec{\Theta})}\right] && (10)\\
&= -\mathbb{E}_{Q(\vec{\Theta})}\left[\log p(x)\right] - \mathbb{E}_{Q(\vec{\Theta})}\left[\log \frac{p(\vec{\Theta}|x)}{Q(\vec{\Theta})}\right] && (11)\\
&= D_{KL}[Q(\vec{\Theta})||P(\vec{\Theta}|x)] - \log p(x) && (12)
\end{aligned}$$

The Kullback-Leibler divergence signifies the information loss inherent in approximating the actual posterior probability distribution using the variational distribution (also called recognition density; Friston, 2010). This positive difference makes free-energy an upper bound on surprise. Importantly, this conceptualization highlights that biological systems striving to minimize free-energy are incentivized to refine their model to accurately approximate the posterior probability distribution. One can see this by considering that, under constant surprise, a closer alignment between the recognition density and the true posterior leads to a reduced Kullback-Leibler divergence and, consequently, smaller free-energy. The Bayesian nature of free-energy minimization is underscored by the fact that as the variational distribution approaches the true posterior, it effectively becomes an approximate posterior probability.

2.3 Active Inference

A central process theory in the free-energy framework is active inference. At the core of active inference sits the idea of a self-sufficient recursion. To explain this idea, it makes sense to introduce the idea of Markov blankets from which active inference inherits. The concept of Markov blankets gives meaning to the term *system* by statistically defining what differentiates it from its environment (Kirchhoff et al., 2018). Markov blankets partition a system into internal, external and blanket states. Blanket states can be differentiated into active and sensory states. Internal and external states are conditionally independent, given the blanket states - meaning they can only influence each other via blanket states. External states influence sensory states, which influence internal states (but are not themselves influenced by internal states) and internal states influence active states, which influence external states (but are not themselves influenced by external states; Kirchhoff et al., 2018). This implies that knowing an internal state is not enough to infer an external state and vice versa.[4]

As argued above, the generative model (internal states) of agents is geared toward the belief that they will minimize free-energy, because policies with lower free-energy are more likely (less surprising). Action in this framework serves as a blanket state and bridges the generative model and the data generating process in that action is sampled from posterior beliefs and influences the external world (Friston et al., 2015). The differentiation between the subjective generative model and the generative process, or data generating process, is important and corresponds to internal and external states from the Markov blanket framework. The generative model is the agent's model of how observations are generated. The data generating process, or generative process, is the actual process happening in the world. Hence, although it is the goal to let the generative model approximate the generative process as closely as possible, optimization involving the generative model does not guarantee that these computations are optimal in any way in the outside world. However, action and perception both serve to minimize the same quantity: variational free-energy. And as we have seen from the free-energy formulation above, to minimize free-energy the generative model and the true posterior should be similar. The self-sufficient recursion is due to the fact that inferring a state of the world in some way makes this perception more likely to be true, since the agent's generative model driving perception is linked to the generative process (the outside world) via action.

Thus far we have said that the agent has a model of the world and infers states of the world while being biased to predict states of the world that minimize free-energy and the free-energy minimizing predictions are then realized via action, which is inferred from posterior beliefs (Friston et al., 2017). The generative model hence manipulates the generative process via action and has to infer the consequences of these manipulations via sensory states, starting a new iteration of this self-fulfilling prophecy loop. Therefore, considering that this process serves to minimize the lower bound of surprise or model evidence, this dynamic can be described by saying that in order to survive, a system is constantly self-evidencing, which is what the concept of active inference alludes to (Kirchhoff et al., 2018).[5]

[4] It has to be noted, however, that the usage of the concept of Markov Blankets in the active inference literature has been criticised as idiosyncratic (see Bruineberg et al., 2022).

[5] This is because minimizing surprise is the same as maximizing self-evidence (Friston, 2010).

3 The Exploration-Exploitation Dilemma, Expected Free-Energy and Multi-Armed Bandit Tasks

3.1 The Exploration-Exploitation Dilemma

So far, we have discussed the plausibility of Bayesian inference in perception and derived that in order to implement Bayesian perception, biological agents should strive to minimize a lower bound on surprise, which is called variational free-energy. This leads to a process theory called active inference, in which systems can be described as being in a self-fulfilling prophecy loop of self-evidencing or minimizing entropy. In other words, we have so far described the learning rule and the decision rule of an agent. The learning rule is Bayesian inference and it defines how new information is integrated into the generative model. The choice rule is active inference and it defines how an agent chooses among alternative options, namely it takes the one that minimizes free-energy.

However, one component is still missing to ensure efficient navigation in complex and ever-changing worlds: *planning*. From the perspective of reinforcement learning, when navigating any kind of sufficiently complex environment, an agent has to weigh different options against each other (Sutton and Barto, 2018). Usually, the ultimate goal of such an agent is to maximize some sort of value. However, the question at any given time is whether the agent should follow a value-maximizing policy or whether it should forage, or explore to find a potentially even better policy. A noteworthy characteristic of many situations involving uncertainty is that one does not know what would have happened if one had acted otherwise (Speekenbrink, 2022). An agent thus faces the so-called exploration-exploitation dilemma. Exploiting carries the risk of missing out discovering better actions while exploring carries the risk of high opportunity costs stemming from not pursuing a known valuable action.

3.2 Expected Free-Energy

To accomodate planning, free-energy should not only be minimized with respect to past and present states, but also long-term - with respect to future states. Hence, to ensure its survival an agent has to minimize its *expected* free-energy. The way in which active inference accounts for this is best explained by closely looking at the involved formalisms. One way to express active inference in context of planning, or expected free-energy is in terms of extrinsic and intrinsic value (Marković et al., 2021;

Parr et al., 2022):

$$G(a) = \underbrace{-\mathbb{E}_{Q(o_t|a_t=a)}[\ln P(o_t)]}_{\text{Extrinsic value}} - \underbrace{\mathbb{E}_{Q(o_t|a_t=a)}\left[D_{KL}\left(Q(\vec{\Theta}|o_t, a_t = a)||Q(\vec{\Theta})\right)\right]}_{\text{Intrinsic value/Novelty}} \tag{13}$$

where o_t stands for observations at time t, a_t for action at time t and $\vec{\Theta}$ for the parameters. One of the achievements of this formulation is that information gain (intrinsic value) and the value of preferred observations (extrinsic value) are expressed in the same metric, namely probability, or nats (Friston et al., 2015; Parr et al., 2022).[6]

When minimizing expected free-energy, the relative contribution of intrinsic and extrinsic value determines whether behaviour will be mainly explorative or exploitative respectively. Optimizing these terms balances exploration and exploitation because, for example, an agent only explores the environment if the information gain/uncertainty reduction afforded by that action trumps the utility/extrinsic value of pursuing alternative actions (Friston et al., 2015). Furthermore, if there are several possible actions that would yield the same utility - maybe because they manifest the same state of the world - then that action is chosen which additionally yields the most information gain. Hence, if it promises to be informative, you manifest the state of the world in the way that is least known to you. We can see an apparent conflict here: the option that yields the most information gain is at the same time the option with the highest entropy, but as stated throughout this thesis, the goal of the agent is to minimize entropy. To resolve this apparent conflict, one has to differentiate between different optimization time-scales. To gain an understanding of the environment and hence to be able to most efficiently reduce entropy in the long-term sometimes requires maximizing entropy in the short-term as only this enables new insights (Friston et al., 2015). This implies that exploration in an active inference agent is not random. Exploration is goal-oriented in the way that those options are explored that promise the highest information gain. In the context of complex environments with large state spaces, which most real-world decision problems are situated in, this makes sense: completely random exploration would introduce redundant behaviour and hinder the exploration of all valuable options (Millidge, 2021).

To sum up, in situations of planning, where epistemic uncertainty (accuracy of one's beliefs about the world) is involved and we can improve the chance of maximizing our goal by learning, uncertainty should play a role in choice behaviour (Speekenbrink, 2022).

[6] Recall that this is possible because in the free-energy framework agents choose actions according to the degree to which they reduce uncertainty, which determines their value.

3.3 Multi-Armed Bandit Tasks

Multi-armed bandits are a classic experimental paradigm to study sequential interaction with an environment that provides feedback (Gershman, 2018; Marković et al., 2021; Speekenbrink, 2022). Each round, an agent chooses an option (one arm from the multi-armed bandit) and receives a corresponding feedback according to the reward distribution of the respective arm. The reward distribution underlying the arms can either remain constant (stationary bandit) or vary over time (restless bandit). While stationary bandits are a classic machine learning problem, we focus on the restless bandit problem as it resembles real-world scenarios more accurately and is hence more relevant for the discussion of behavioural fit of the algorithms tested in this thesis (Marković et al., 2021).

The bandits considered in this thesis follow a normal distribution with a mean that remains constant for a given number of trials and then changes and a constant standard deviation throughout the whole experiment. This makes the exploratory behaviour required more complex than in the stationary bandit case, because due to the changing distribution some degree of exploration is required throughout the experiment (as opposed to the stationary bandit, where exploration can be front loaded; Marković et al., 2021; Speekenbrink, 2022).

A classic illustration of why restless bandits are a decision and learning scenario that resembles real-world situations is the dinner choice in a restaurant. Even if we have been to the same restaurant before and have some degree of knowledge about the menu, it is possible that the chef has changed or has a bad day or that the owners have changed and so on. Hence, the choice of the menu item for the evening should take the inherent stochasticity of the restaurant into account, even if one loved the steak one had on the last visit. Consider again the information bottleneck mentioned above: after having chosen one item, it is usually not possible to know how the alternatives would have tasted. Hence, in order to know whether one likes the fish (maybe even better than the steak), one has to try it out at some point.

4 Methods

4.1 The Two-Armed Bandit Task

To investigate whether active inference algorithms are behaviourally plausible in relevant scenarios which involve uncertainty, feedback from the environment, planning and learning, we chose to conduct experiment 2, a restless two-armed bandit task, from Gershman (2018). In this experiment, 44 participants played 200 trials seperated into 20 two-armed bandits in blocks of 10 trials. After each block the mean reward for both arms was drawn from a normal distribution with mean 0 and variance 100. The reward on each trial for the respective arm was then drawn from a normal distribution with mean $\mu_k = k$ with k indicating the selected arm and variance 10 (Gershman, 2018).

Hence, we consider the K multi armed bandit task parametrised by the vector of means $\mathbf{m} = [\mu_1, \mu_2, ..., \mu_K]$ and $\mathbf{s} = [\sigma_1, \sigma_2, ..., \sigma_K]$ where each $k \in \{1, ..., K\}$ follows a normal distribution $\mathcal{N}(\mu_k, \sigma_k)$ and is associated with an *i.i.d* random variable o_t at time $t \in \{1, ..., T\}$. The generative model at time t is

$$p(o_{1:T}, \mathbf{m}, \mathbf{s}|a_{1:T}) = p(\mathbf{m}|a_t)p(\mathbf{s}|a_t)\prod_{t=1}^{T} p(o_t|\mathbf{m}, \mathbf{s}, a_t) \tag{14}$$

with $p(\mathbf{m}|a_t)$ and $p(\mathbf{s}|a_t)$ being the posterior beliefs about the mean and sd of the chosen arm and

$$p(o_t|\mathbf{m}, \mathbf{s}, a_t) = \prod_{k=1}^{K} \left[p(o_{t,k}|\mu_k, \sigma_k)\right]^{\mathrm{I}_{k=a_t}} \tag{15}$$

with $\mathrm{I}_{k=a_t}$ being the indicator function, which is 1 if the action denotes the respective arm and 0 otherwise and (14) being the observation likelihood of the generative model. In order to discuss how the posterior beliefs are computed, we next introduce the shared Bayesian Inference learning rule.

4.2 The Bayesian Inference Learning Rule

As indicated above, we used the same learning rule for all choice algorithms, namely Bayesian inference. Specifically, we used the Bayesian update rule with unknown mean μ and unknown precision (inverse variance) $\lambda = \frac{1}{\sigma^2}$. The Normal-Gamma distribution is the appropriate conjugate prior for the normal distribution with unknown mean and unknown precision (Bishop and Nasrabadi, 2006).

The prior hyperparameters are α: the shape parameter of the Gamma distribution, β: the scale parameter of the Gamma distribution, μ_0: the mean of the normal distribution in the prior and λ_0: the precision of the normal distribution in the prior. After observing data, the updated hyperparameters are given by:

$$\alpha' = \alpha + \frac{n}{2} \tag{16}$$

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\lambda}{2\lambda + n} \cdot (\bar{x} - \mu_0)^2 \tag{17}$$

$$\mu' = \frac{\lambda\mu_0 + n\bar{x}}{\lambda + n} \tag{18}$$

$$\lambda' = \lambda + n \tag{19}$$

Where: n is the number of observed data points, $\bar{x}$ is the sample mean of the observed data, α' and β' are the updated shape and scale parameters of the Gamma distribution, μ' is the updated mean of the normal distribution and λ' is the updated precision (inverse variance) scaling factor (Bishop and Nasrabadi, 2006). To avoid clutter, we omitted the subscript k for all of these parameters, indicating that the learning process happens per arm. These formulas define how the hyperparameters of the Normal-Gamma distribution are updated after observing new data. They capture the combination of prior information and data likelihood to obtain the updated posterior distribution for the unknown mean and precision.

However, the Bayesian inference process presented thus far has the pitfall that it is tuned to a stationary environment. To adapt for this, we use the exponential weighting sliding window approach in which we can adapt how much weight we give previous datapoints in the running mean estimate (Bodenham, 2012). The algorithm calculates a set of weights and subsequently applies these weights recursively to the data samples. As the temporal distance from the present increases, the weighting factor diminishes exponentially, but never reaches zero. In essence, this implies that the statistical influence of recent data on the current sample outweighs that of older data. The idea behind this is that in a dynamically changing environment it makes sense to forget about previous data samples as they provide limited information when the underlying distribution has changed. The value of the memory parameter influences the weighting factors' rate of change - with a higher memory giving more weight to older data points and a lower forgetting factor giving more weight to more recent data (Bodenham, 2012). A memory of 1 means perfect memory with all sampled data points being weighted equally.

$$w_{i,\varphi} = \varphi w_{i-1} + 1 \tag{20}$$

$$\bar{x}_{i\varphi} = (1 - \frac{1}{w_{i\varphi}})\bar{x}_{i-1,\varphi} + (\frac{1}{w_{i\varphi}})x_i \tag{21}$$

with φ being the memory parameter, $w_{i,\varphi}$ being the weighting factor applied to the current data sample, x_i the current data input sample and $\bar{x}_{i\varphi}$ the moving average at the current sample.

Integrating the sliding window approach into the Bayesian Inference rules above, we get

$$\alpha'' = \alpha + \frac{n}{2} \tag{22}$$

$$\beta'' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x}_{i\varphi})^2 + \frac{n\lambda}{2\lambda + n} \cdot (\bar{x}_{i\varphi} - \mu_0)^2 \tag{23}$$

$$\mu'' = \frac{\lambda\mu_0 + n\bar{x}_{i\varphi}}{\lambda + n} \tag{24}$$

$$\lambda'' = \lambda + n \tag{25}$$

with the value of $\bar{x}_{i\varphi}$ being dependent on the memory parameter φ.

4.3 Action Selection

4.3.1 Active Inference

Above in (13) we have already seen an active inference formulation in the context of planning. This formula was centered around the notions of extrinsic value (preferred options) and intrinsic value (information gain) and served to illustrate how active inference dissolves the exploration-exploitation dilemma parsimoniously - that is, without the addition of ad hoc parameters regulating the exploration behaviour (Friston et al., 2015). For the active inference choice rule presented here, we used a reformulation of this (Marković et al., 2021; Parr et al., 2022)

$$G(a) = \underbrace{D_{KL}(Q(o_t|a_t = a)||P(o_t))}_{\text{Risk}} + \underbrace{\mathbb{E}_{Q(\vec{\Theta})}\Big[H[(o_t|\vec{\Theta}, a_t = a)]\Big]}_{\text{Ambiguity}} \tag{26}$$

Where the first term is called risk, because it is acknowledged that there is some uncertainty in the system that needs to be accounted for. This means that even if we have a good idea of the underlying distribution, there is no deterministic mapping from action to a desired outcome as there is some stochasticity inherent to the system (Speekenbrink, 2022). Taking this into account, the risk term ensures that actions are chosen whose probabilistic outcomes (encoded by the predictive distribution $Q(o_t|a_t = a)$) match the preferred outcome (encoded by $P(o_t)$) in the sense of a KL-divergence (Parr et al., 2022).

The second term is called ambiguity, as it encodes the expected inaccuracy due to an ambiguous mapping between states and outcomes (Parr et al., 2022). What this means is that even if we have confident beliefs about the states generating outcomes, the distribution of anticipated outcomes might have a high variance (high entropy). As such, if the expected entropy of the distribution is high, this limits the information to be gained about the underlying parameters by observing an outcome (Parr et al., 2022).

As this is just a reformulation of (13) it should not be surprising that active inference agents minimizing expected free-energy according to (26) balance the exploration-exploitation dilemma in the same way. Out of two options that are equally likely to result in the preferred outcome (same risk) the one is chosen that at the same time yields the highest information gain (least ambiguity).

Let us first look at the term $D_{KL}(Q(o_t|a_t = a)||P(o_t))$, where the predictive distribution $Q(o_t|a_t = a)$ is obtained by integrating over the model parameters involved in our learning rule, μ and λ.

$$Q(o_t|a_t = a) = \int Q(o_t|a_t = a, \mu, \lambda)Q(\mu, \lambda|a_t = a)d\lambda d\mu \tag{27}$$

and according to the assumptions of a normal distribution with unknown mean and variance

$$Q(o_t|a_t = a, \mu, \lambda) = \mathcal{N}(o_t|\mu, \lambda^{-1}) \tag{28}$$

$$Q(\mu, \lambda|a_t = a) = \text{Normal-Gamma}(\mu, \lambda|\mu_0(a), \lambda_0(a), \alpha(a), \beta(a)) \tag{29}$$

where $\mu_0(a), \lambda_0(a), \alpha(a), \beta(a)$ are the hyperparameters that are to be updated in the learning rule. It can be shown that in this context of Bayesian inference involving a normal distribution with unknown mean and variance the posterior predictive $Q(o_t|a_t = a)$ follows a t-distribution (Bishop and Nasrabadi, 2006).

$$Q(o_t|a_t = a) = T(o_t|\tilde{\lambda}(a), \tilde{\mu}(a), \tilde{\nu}(a)) \tag{30}$$

$$\text{where} \quad (\tilde{\lambda}(a), \tilde{\mu}(a), \tilde{\nu}(a)) = f(\mu_0(a), \lambda_0(a), \alpha(a), \beta(a)) \tag{31}$$

For further details on the derivation of (29) and (30), see Appendix A.

$P(o_t)$ denotes preferences over outcomes. In line with previous research on multi-armed bandits, we chose the Boltzmann exponential weighting scheme to encode such preferences (Cesa-Bianchi et al., 2017). Hence, P(o) = exp(ηo), where η is a parameter controlling the exploration-exploitation trade-off. The role of this Boltzmann exploration scheme is to exert an influence on the choice behaviour that is proportional to its estimated utility (Cesa-Bianchi et al., 2017). This implementation makes the Kullback-Leibler divergence analytically solvable and favors higher outcomes (hence, functions as preference). Therefore, the Kullback-Leibler divergence is given by:

$$\begin{aligned} KL(Q(o_t|a_t = a)||P(o_t)) &= \mathbb{E}_{Q(o_t|a_t=a)}[\eta o_t] - H[Q(o_t|a_t = a)] && (32) \\ &= \eta\tilde{\mu}(a) - H[T(\tilde{\lambda}(a), \tilde{\mu}(a), \tilde{\nu}(a))] && (33) \end{aligned}$$

with the entropy of a t-distribution, $H[T(\tilde{\lambda}(a), \tilde{\mu}(a), \tilde{\nu}(a))]$ being

$$\frac{\tilde{\nu}(a)+1}{2}\left[\psi\left(\frac{1+\tilde{\nu}(a)}{2}\right) - \psi\left(\frac{\tilde{\nu}(a)}{2}\right)\right] + \ln\left[\sqrt{\tilde{\nu}(a)}B\left(\frac{\tilde{\nu}(a)}{2}, \frac{1}{2}\right)\right] \tag{34}$$

where ψ is the digamma function and B is the beta function. Putting this together, the Kullback-Leibler divergence can be analytically solved according to

$$KL(Q(o_t|a_t = a)||P(o_t)) = \eta\tilde{\mu}(a) - \frac{\tilde{\nu}(a)+1}{2}\left[\psi\left(\frac{1+\tilde{\nu}(a)}{2}\right) - \psi\left(\frac{\tilde{\nu}(a)}{2}\right)\right] + \ln\left[\sqrt{\tilde{\nu}(a)}B\left(\frac{\tilde{\nu}(a)}{2}, \frac{1}{2}\right)\right] \tag{35}$$

Let us now take a look at the expectation of the conditional entropy of the observation likelihood $\mathbb{E}_{Q(\vec{\Theta})}\left[H[(o_t|\vec{\Theta}, a_t = a)]\right]$, which we add to the Kullback-Leibler divergence in order to calculate the expected free-energy of a given action. As we have already defined the observation likelihood, we just need to fill it in here:

$$\begin{aligned}
\mathbb{E}_{Q(\mu,\lambda|a_t=a)}\left[H[\mathcal{N}(o_t|\mu,\lambda^{-1})]\right] &= \mathbb{E}_{Q(\mu,\lambda|a_t=a)}\left[\frac{1}{2}\ln(2\pi e/\lambda)\right] && (36)\\
&= \frac{1}{2}\ln(2\pi e) - \mathbb{E}_{Q(\lambda|a_t=a)}\left[\ln(\lambda)\right] && (37)\\
&= \frac{1}{2}\ln(2\pi e) - \mathbb{E}[\ln(\lambda)|\lambda \sim \text{Gamma}(\alpha(a),\beta(a))] && (38)\\
&= \frac{1}{2}\ln(2\pi e) + \log(\beta(a)) + \psi(\alpha(a)) && (39)
\end{aligned}$$

where we first pull the constants ($\frac{1}{2}\ln(2\pi e)$) out of the expectation since they don't depend on the distribution. Equation (38) makes explicit that we take the expectation of $\ln(\lambda)$ with respect to its distribution, which is the gamma distribution - parametrised by $\alpha(a)$ and $\beta(a)$. The expectation of the logarithm of a gamma distributed variable can be worked out to be $\log(\beta(a)) + \psi(\alpha(a))$ (Soch et al., 2022).

Putting all of this together, the expected free-energy of each action in the context of Bayesian inference with an assumed normal distribution with unknown mean and variance is:

$$\begin{aligned}
G(a) = \eta\tilde{\mu}(a) - \frac{\tilde{\nu}(a)+1}{2}\left[\psi\left(\frac{1+\tilde{\nu}(a)}{2}\right) - \psi\left(\frac{\tilde{\nu}(a)}{2}\right)\right] + \ln\left[\sqrt{\tilde{\nu}(a)}B\left(\frac{\tilde{\nu}(a)}{2},\frac{1}{2}\right)\right] \quad (40)\\
+\frac{1}{2}\ln(2\pi e) + \log(\beta(a)) + \psi(\alpha(a))
\end{aligned}$$

The last part of this choice rule is the softmax transformation (Sutton and Barto, 2018). According to the softmax rule, the probability of choosing option j is

$$P(C_{max} = j) = \frac{e^{G(j)}}{\sum_{i=1}^{N} e^{G(i)}} \quad (41)$$

However, because we want to choose the option with the least expected free-energy, we use what could be called a softmin choice rule, according to

$$P(C_{min} = j) = 1 - P(C_{max} = j) \quad (42)$$

This softmin rule takes as input the expected free-energy of all options at any given time and outputs the relative goodness of these choices in terms of probabilities. The agent then randomly chooses between the options with their respective probabilities.

4.3.2 Upper Confidence Bound

The Upper Confidence Bound (UCB) choice rule is often used as a standard to test reinforcement learning algorithms and it implements a form of optimism in the face of uncertainty (Gershman, 2018; Marković et al., 2021). This means that between options with the same expected reward, the option is chosen that has highest uncertainty. The idea is that high estimation uncertainty is an indicator of possible information gain via exploring this option (Gershman, 2018). In that endeavour, we calculate confidence intervals around the expected rewards and choose the option with the highest UCB. Here, we compute the choice rule according to UCB as follows

$$P\left(C=j\right)=\begin{cases}1 & \text{if },\forall k\neq j,\bar{x}_{j\varphi}+\beta(\sqrt{\frac{1}{\lambda_j}})>\bar{x}_{k\varphi}+\beta(\sqrt{\frac{1}{\lambda_k}})\\ 0 & \text{otherwise}\end{cases}\tag{43}$$

where β is a hyperparameter and determines the upper limit of the confidence interval. However, for reasons of comparison, this deterministic choice rule is turned into a probabilistic choice rule by using the softmax transformation. Hence

$$P(C=j)=\frac{e^{ucb(j)}}{\sum_{k=1}^{N}e^{ucb(k)}}\tag{44}$$

where $ucb_i=\bar{x}_{i\varphi}+\beta(\sqrt{\frac{1}{\lambda_i}})$.

4.3.3 Thompson Sampling

The second choice strategy that we will compare the active inference agent to is Thompson sampling (Thompson, 1933). In Thompson sampling, the algorithm samples a reward from each arm's prior distribution and selects the arm with the highest sampled value. This process is inherently stochastic due to its random sampling aspect. Hence, we choose according to

$$P\left(C=j\right)=\begin{cases}1 & \text{if }\tilde{m}_j=\max_k\left(\tilde{m}_k\right)\\ 0 & \text{otherwise}\end{cases}\qquad \tilde{m}_j\sim\mathcal{N}\left(m_j,\frac{\sigma_j^2}{\sqrt{\lambda_j}}\right)\tag{45}$$

where $\tilde{m}_j$ denotes a sample from the prior distribution of mean rewards of the respective arm. The expression $\frac{\sigma_j^2}{\sqrt{\lambda_j}}$ is used in place of the standard deviation to include a measure of confidence in the estimate.

4.3.4 Random Agent

Lastly, we also considered the performance of a random agent. The random choice rule embodies a stochastic process where the agent generates $k \in \{1, ..., K\}$ uniform random variables in the interval [0,1], where K is the number of arms. The agent then uses the normalized probabilities as choice probabilities of the respective arm. The choice rule thus takes the following form

$$P(C = j) = \frac{u_j}{\sum_{k=1}^{K} u_k} \tag{46}$$

where $u_k \sim U(0, 1)$. This way, the agent disregards any learning and chooses completely at random.

The random choice rule serves as a baseline that does not take uncertainty into account. Consider again that in active inference, UCB and Thompson sampling the exploration of uncertain options is balanced with the exploitation of known valuable options, albeit in a different way in each algorithm. The aim of this thesis is to find out whether active inference provides a better fit to model human exploration and exploitation behaviour than classic and established algorithms for which we chose UCB and Thompson sampling as representatives.

4.4 Participants

We are conducting experiment 2 from Gershman (2018), where 44 participants were recruited via Amazon Mechanical Turk and were paid \$ 1.50 for their participation. With the structure of the task having been outlined in 4.1, we shall focus on the instructions that the participants received here. The participants were instructed to choose the arm that maximizes their total reward. The exact instructions that Gershman (2018, p. 36) provided read as follows:

> In this task, you have a choice between two slot machines, represented by colored buttons. When you click one of the buttons, you will win or lose points. Choosing the same slot machine will not always give you the same points, but one slot machine is always better than the other. Your goal is to choose the slot machine that will give you the most points. After making your choice, you will receive feedback about the outcome. You will play 20 games, each with a different pair of slot machines. Each game will consist of 10 trials.

Hence, participants were explicitly informed about the stochastic nature of the experiment environment.

4.5 Statistical Analyses

To investigate the hypothesis of this thesis, that active inference models human performance better than alternative benchmark algorithms, we chose maximum likelihood estimation. Utilizing the SciPy package optimize.minimize, for each participant we searched for each choice rule those parameter values that maximized the likelihood of the choices the participant actually made (Virtanen et al., 2020). Specifically, we used the Powell algorithm, which iteratively changes the respective parameter values and searches for local minima, taking the minimum of these local minima to be the global minimum (Powell, 1964). Note that in order to accommodate the fact that the Powell algorithm searches for minima instead of maxima, we computed the negative log-likelihood, transforming the maximization problem into a minimization problem. As likelihood function, we used (14) for all tested models.

The parameters that we optimized were eta (η) and memory (φ) in case of the active inference agent, beta (β) and memory (φ) in case of the UCB algorithm and φ in case of Thompson sampling. We initialised the Powell algorithm with arbitrarily chosen parameter values, as this algorithm was specifically chosen, because it samples over the whole parameter space. Very roughly, η determines the exploration behaviour (with higher values leading to less exploration), φ determines the memory of the agent (with higher values meaning better memory, i.e. considering more distant data points) and β determines the upper limit of the confidence interval that is added to the posterior mean in the UCB choice rule. We specified that the minimization algorithm searches in the parameter space of [0, 1] for both η and φ, while the β bounds were [0, 2.5758], as $\beta \approx 2.5758$ provides the upper limit of the 99% confidence interval.

To test whether one of the models fits the data significantly better than the other models, we conducted pairwise comparisons of the relative likelihood of all models, using the Akaike information criterion (AIC) statistic (Wagenmakers and Farrell, 2004). The tests were conducted according to the following procedure:

$$AIC_1 = 2 * x_1 + 2 * q_1 \tag{47}$$

$$AIC_2 = 2 * x_2 + 2 * q_2 \tag{48}$$

$$AIC_{min} = min(AIC_1, AIC_2) \tag{49}$$

$$B_i = exp((AIC_{min} - AIC_i)/2) \tag{50}$$

$$p_i = B_i/sum(B_i) \tag{51}$$

where x_i is the negative log likelihood of model i and q_i is the number of parameters of model i. If a model has a relative likelihood of $p_i > 0.95$, it can be concluded that it fits the data significantly better. This procedure is to be preferred over just choosing the model that has a lower AIC statistic, because if the AIC difference is very small, implying that there might be no practical difference between the models, the present method does not judge one to be better than the other (Wagenmakers and Farrell, 2004).

The performance of the models was visually compared to the optimal strategy. The optimal choice was taken to be the arm that had the highest true reward distribution mean on the given trial (Marković et al., 2021).

5 Results

5.1 Descriptive Statistics

Figure 1 depicts an exemplary illustration of the stochasticity of the environment. It shows how the reward distributions of both arms change over the course of the experiment for a random participant (here participant 26). As outlined in 4.1, each reward distribution mean shown in this plot has been drawn from a normal distribution with mean 0 and variance 100. Note that these values are not directly accessible to the participant, but they depict distribution means from which individual rewards that the agent receives are drawn on each trial.

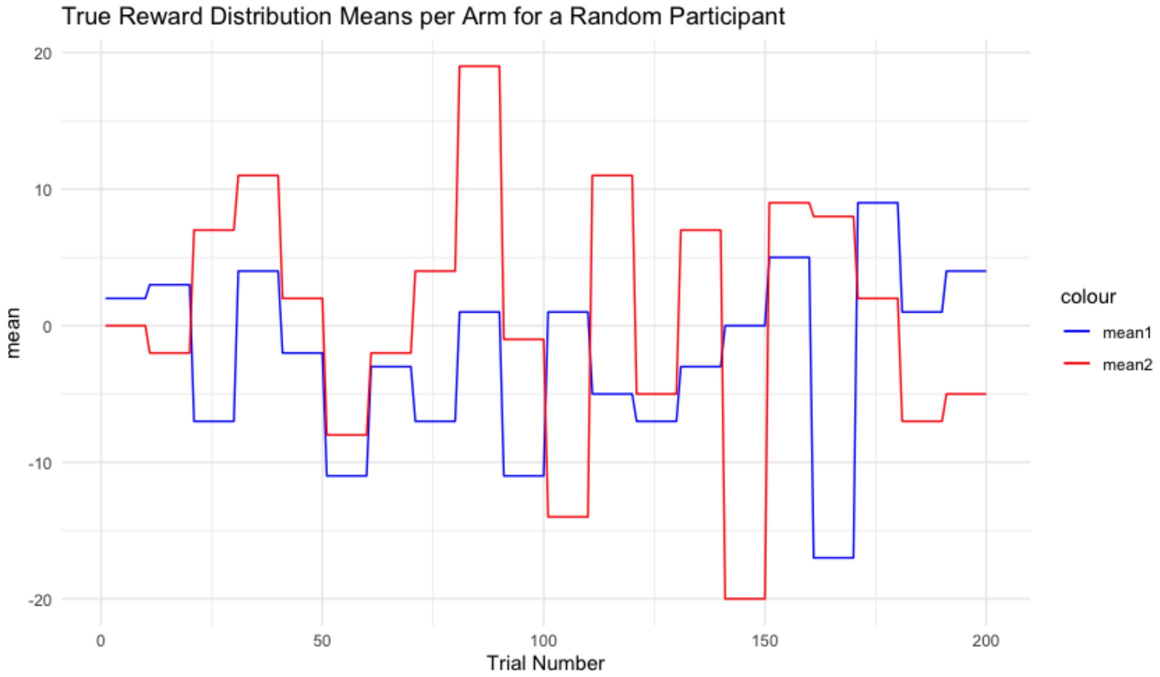

Figure 1: This plot shows for both arms the distribution means from which the respective rewards are drawn for a randomly chosen participant 26. The blue line plot depicts the changing means of arm 1 and the red line plot depicts the changing means of arm 2. A new mean is drawn every 10 trials.

One way to preliminarily compare the algorithms and humans is to investigate the received reward, since maximizing reward is ultimately the goal of this experimental design. However, as all algorithms are parameterised, the question arises what parameter values should be chosen for the comparison. As the goal of this thesis was to investigate whether active inference models human data better than other benchmark algorithms, a meaningful comparison in this context is to use those parameters that the maximum likelihood estimation (outlined in section 4.5) determined. The results of the maximum likelihood analysis detailing the parameter value distributions can be seen in the inferential statistics section below. Here, we used these parameter values for each algorithm and each participant to simulate how the algorithm fitted with the respective optimal parameters per participant would have solved the task. Note that this is not the same as running the optimization algorithm. In the optimization algorithm, we estimate the likelihood of choices made by humans, given the respective model. Here, we use the parameters that this procedure outputs to let the models solve the two-armed bandit task themselves.

Importantly, while the algorithms are initialised with the individually fitted hyperparameters of the choice rules per participant, each simulated participant starts with the same Bayesian priors. These prior expectations are as follows: a mean of $\mu = 0$, a precision of $\lambda = 1$, an alpha of $\alpha = 1$ and a beta of $\beta = 2$. The variance priors α and β, which parameterise the inverse gamma distribution of the variance of the assumed normal distribution with unknown mean and variance have been chosen as such, because they centre the probability density function around smaller values while also allowing for higher values, essentially reflecting that the participants are uninformed about the underlying distribution and assume that the mean and variance are in the same order of magnitude - which we deemed a realistic assumption. The prior mean μ has been chosen to be 0, because the participants were not informed about the underlying distribution being positive or negative and should therefore be agnostic. Furthermore, the precision parameter has been chosen to be $\lambda = 1$ to model a low degree of confidence in the estimate of the mean.

Firstly, we consider the performance of the human participants, which received a mean reward of $M = 3.67$, $SD = 7.63$ and a cumulative reward of 32268. The optimal strategy received a mean reward of $M = 4.77$, $SD = 7.62$ and a cumulative reward of 41966.56 and the random strategy received a mean reward of $M = 0.41$, $SD = 8.68$ and a cumulative reward of 3614.16. The active inference agent received a mean reward of $M = 1.14$, $SD = 8.54$ and a cumulative reward of 10027.71, while the Thompson sampling algorithm achieved a reward distribution of $M = 0.96$, $SD = 8.85$ and a cumulative reward of 8472.61. Lastly, the UCB algorithm received a mean reward of $M = 0.76$, $SD = 8.78$ and a cumulative reward of 6679.35. Thus, in terms of ranking the algorithms and human participants based on the received cumulative reward, the optimal strategy seems to be most successful, followed by human participants, active inference, Thompson sampling, UCB and the random agent in that order.

However, since cumulative reward provides only limited information in terms of how well the algorithms model human data, it is helpful to plot the performance of the algorithms and humans over the course of the experiment. In Figure 2, we chose to plot the average performance of each algorithm over the course of the average experiment. This means that we are plotting the mean of each algorithm for each trial number (from 1 to 200, as this is the length of the experiment for each participant) for each participant. In this graph our primary interest is not in the intercepts of the curves, but rather in their slopes. This is because, as stated above, the question addressed in this thesis is not whether active inference can match human performance in terms of total reward received, but rather whether it can model it well. An important aspect of the latter question is whether it reacts similarly to the same environmental patterns.

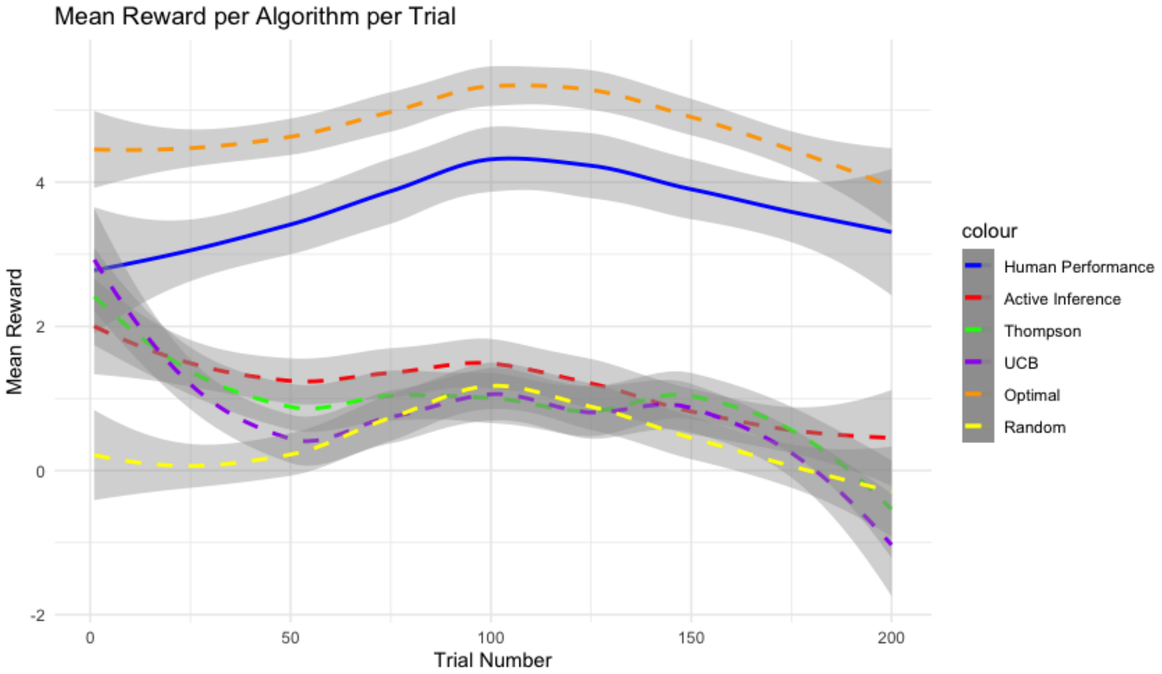

Figure 2: This plot depicts a regression over the mean received reward of all participants for each trial for humans (blue), active inference (red), Thompson sampling (green), UCB (purple), the optimal strategy (orange) and a randomly choosing agent (yellow). The shaded areas depict 95% confidence intervals. Note that the choice of averaging over trials was taken for the purpose of comparison and illustration of the algorithms. Although the structure of this plot might suggest otherwise, the true reward means are drawn every 10 trials throughout the whole experiment. This means that the true reward means are not the same for all participants. Hence, from averaging over each trial of all participants, we can investigate whether the aggregate of the human participants and the initialisations of our algorithms behave in a way that reflects the aggregate environmental dynamic. Every inference we draw about choice patterns and/or learning and planning behaviour thus refers to the macroscopic scale rather than processes on the level of individual humans or algorithm instantiations.

Firstly, we can observe that the shape of the optimal, human and random curves seem to be very similar. They tend to have the same slope in the same x-axis regions, approximating a bell-shaped curve. That this pattern is a meaningful one becomes clear when looking at the mean of the distributions of both arms depicted in Figure 3, which also exhibits this bell-shaped pattern. This comparison is interesting, because the mean of of the distribution of both arms reflects the expected value of the environment if choosing without a specific pattern - that is, randomly. The latter can also be seen when comparing the random agent curve from Figure 2 with the mean of both distributions in Figure 3. They seem almost identical. Therefore, judging based on the reward distribution pattern visible in Figure 3, the shape of the aforementioned curves reflects a property of the environment. Of the three remaining choice rules, active inference approximates the bell-shaped pattern of the environmental reward distribution most. The Thompson and UCB algorithms seem to follow local trends of the respective arms more than humans and active inference do. For example, consider trials 100 until 200. While the reward achieved by humans and active inference seems to follow the overall trend of declining reward due to declining reward distribution means, Thompson sampling and UCB both exhibit a short increase in rewards roughly from trial 130 to 150. This is consistent with the increase in the true reward mean of arm 2 around the same trials. While this is beneficial for a while, it seems like both algorithms follow the trend of this arm further, resulting in a steep reward decline that is avoided by active inference and humans.

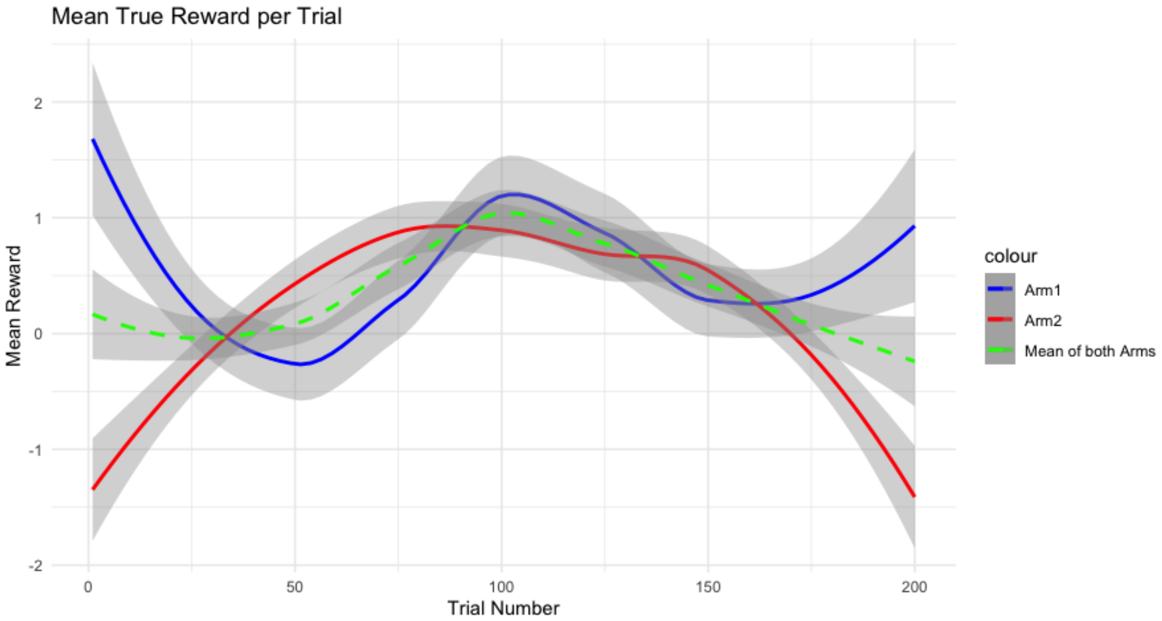

Figure 3: This figure depicts regressions of the mean of the reward distributions of arm 1 (blue) and arm 2 (red) averaged over each trial for all participants and their mean (dashed green line). The shaded areas depict 95% confidence intervals. The mean is interesting in this case because it reflects the expected value of the environment in case of choosing at random.

In order to investigate how individual humans and instantiations of the algorithms react to changes in the environment, it is fruitful to consider their reward received in randomly chosen experiments. In Figure 4, we can see the reward received for each trial for four randomly chosen participants and the algorithms fitted with the optimized parameters for that specific participant. Humans are depicted as dashed lines while the algorithms are solid lines. The overall pattern is similar: the slope of the active inference curve seems to follow the human and optimal curve while Thompson sampling and UCB seem to follow local trends more. However, a notable difference to the aggregate plots is the relationship of the random curve to the optimal, human and active inference curve. It seems like a good sanity check in the aggregate case that active inference has a shape similar to the random curve, as each point in the random curve depicted the mean of several independent draws of normally distributed variables, which, according to the central limit theorem, will itself be normally distributed and therefore reflect a characteristic of the environment (Kwak and Kim, 2017). The desirability of modelling random behaviour over the course of a single experiment, however, is less clear, as the central limit theorem is not as applicable for such a sample size. Instead the opposite seems to be the case: it is undesirable to showcase behaviour that seems completely random. And indeed, based on visual exploration there are more pronounced differences between random behaviour and the behaviour showcased by active inference and humans.

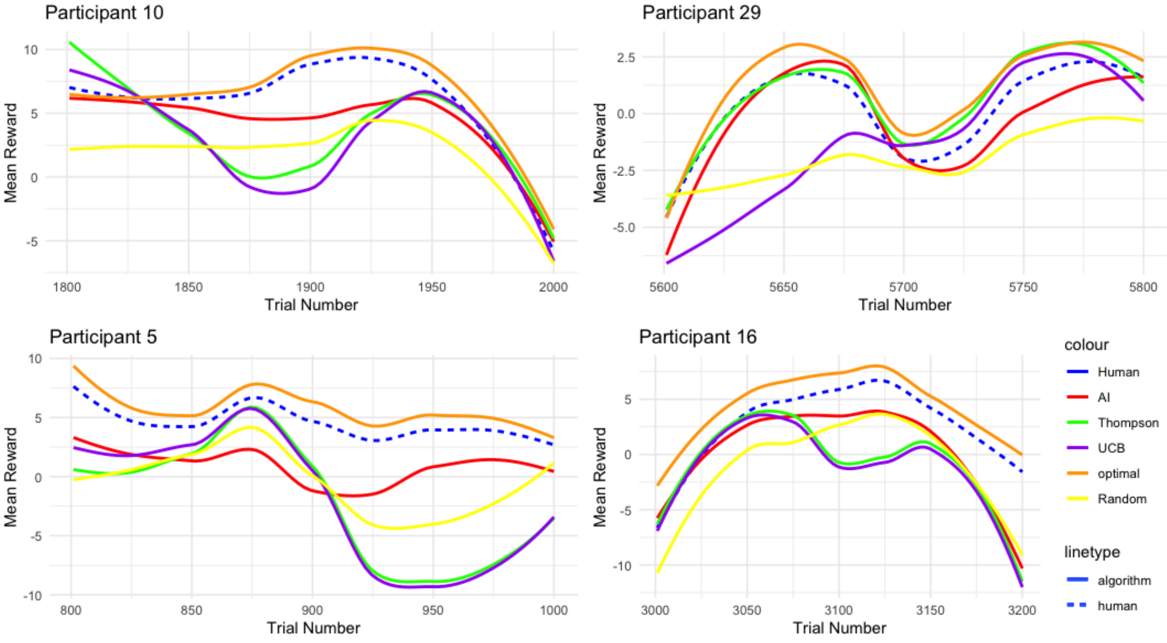

Figure 4: This plot depicts regressions of the mean reward received by four randomly chosen human participants (blue) and the algorithms active inference (red), Thompson sampling (green), UCB (purple), the optimal strategy (orange) and random behaviour (yellow) optimized for modelling that respective participant. The 95% confidence intervals were omitted in this plot for illustrative purposes. Human data is depicted as a dashed line. The trial numbers are based the overall numbers of trials. For example, because each participant did 200 trials, the trial numbers of participant 5 range from 800 (4*200 trials before her) until 1000.

As the centrepiece of this thesis is the active inference algorithm, it seems fruitful to investigate the interplay of the two hyperparameters involved in this choice rule, namely memory (φ) and eta (η). Figure 5 depicts the influence of these two parameters on the cumulative reward received by the agent.

This figure depicts every eta memory combination in the interval [0,1] with step size 0.1. It can be seen that the maximum cumulative reward is achieved with a memory value of 0.1 (recall that lower memory values mean weighting the more recent data points higher than the more temporally distant ones) and an exploration parameter of 0.5. From the free-energy formulation (41) it follows that η is the weighting factor of the posterior mean. Hence, a higher eta indicates a higher role of utility in the decision process and therefore less exploration behaviour. The different line plots belong to different memory values and it is evident that less memory leads to higher cumulative reward. This verifies the exponential sliding window approach as we implemented it based on the assumption that it seems counterproductive to consider all data points in a dynamic environment with constantly changing means. However, since the mean only changes every 10 trials and some learning in that time should be helpful, it makes sense that the maximum reward is achieved with a memory of 0.1 and not 0. Further verification comes from the fact that full memory (meaning all data is weighted equally) performs worse than all other memory values which include some forgetfulness. This holds for all values of eta except for 0. The reason for this may be that when eta is 0, much of the information from the learning rule is lost (as the posterior mean has no influence anymore on the free energy calculation). The high memory value probably compensates for that loss.

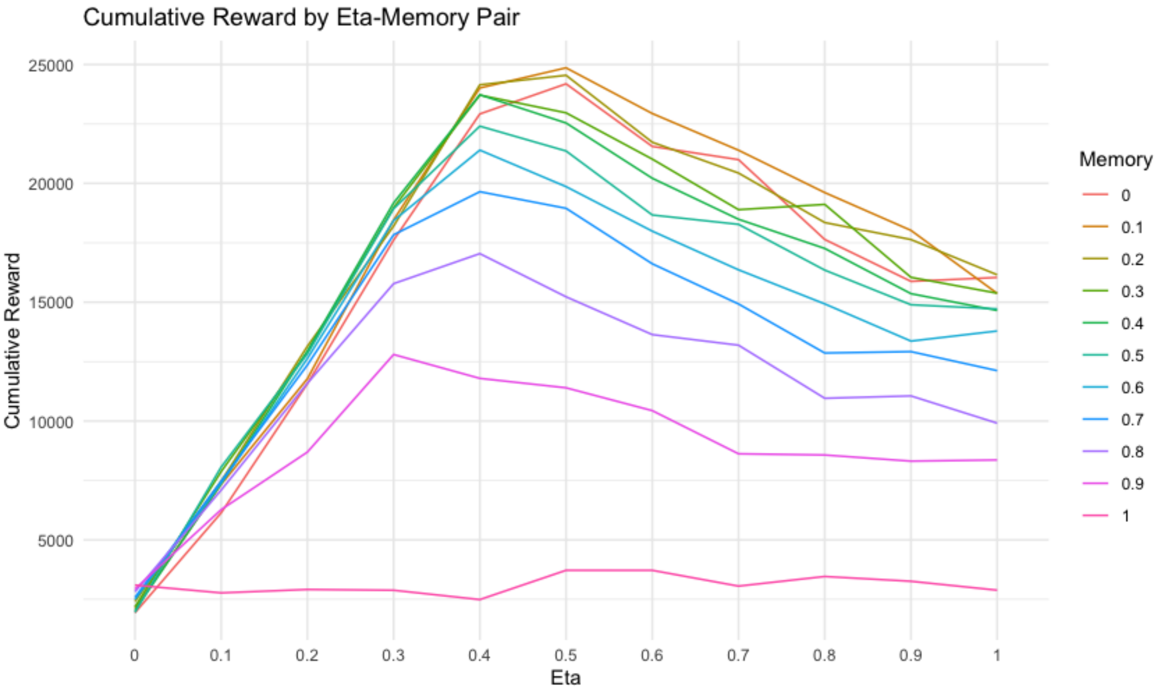

Figure 5: Mean cumulative reward per eta memory pair. A memory factor of 1 (perfect memory) has worst performance for all eta values except for 0.

5.2 Inferential Statistics

The main analysis in this thesis was the a maximum likelihood analysis according to the procedure outlined in section 4.5. To recap, we fed all algorithms the choices that humans made and, using the SciPy package optimize.minimize, searched for those parameter values of the respective algorithm that maximized its likelihood (Virtanen et al., 2020). The results of this analysis are visualised above in Figures 2 and 4. We then compared the relative likelihoods of the different algorithms via the Akaike Information Criterion (AIC). This procedure tests our hypothesis that the free-energy minimizing active inference agent models human data better than classic benchmark algorithms. In order to avoid overcrowded tables with all parameter values for each participant, we chose to report the distribution of the parameters for each algorithm across all participants here.

The parameters of the active inference agent were as follows: eta had a mean of $M = 0.44$, $SD = 0.18$ and the memory parameter had a mean of $M = 0.50$, $SD = 0.22$. The UCB agent beta parameter was distributed with a mean of $M = 2.12$, $SD = 0.54$ while the UCB memory parameter had a mean of $M = 0.60$, $SD = 0.33$. Lastly, the Thompson sampling memory parameter had a mean of $M = 0.45$, $SD = 0.20$ and the random agent memory parameter had a mean of $M = 0.47$, $SD = 0.18$.

Overall, the active inference agent showcased the highest likelihood (lowest AIC) given the human performance of the experiment with an overall mean AIC of $M = 273.35$, $SD = 8.94$, followed by the random agent with a mean AIC of $M = 326.57$, $SD = 9.59$. Thompson sampling showed a mean AIC of $M = 830.13$, $SD = 381$, with UCB showcasing the worst fit (mean AIC of $M = 954.61$, $SD = 464.56$). A table with all likelihood and AIC values can be seen in Appendix B. It is noteworthy that for all participants without exception the likelihood and AIC corresponding to the active inference algorithm are lowest, indicating the best fit of active inference to all participants.

The relative likelihood comparison resulted in the active inference agent having a significantly better fit than the random agent ($p = 0.99$), UCB ($p = 1$) and Thompson sampling ($p = 1$), while the random agent shows a significantly better fit than both Thompson sampling ($p = 1$) and UCB ($p = 1$). Lastly, Thompson sampling shows a significantly better fit than UCB ($p = 1$).[7] This confirms our hypothesis that the active inference algorithm actually models the human data best among the alternatives tested - followed by the random agent, Thompson sampling, and UCB in that order.

To investigate the exploration behaviour of the agents, we calculated the percentage of trials in which the algorithm did not choose the option with the highest estimated posterior mean. Averaged

[7] One might be surprised that p-values above 0.95 are deemed significant here. However, this is due to the way this test is computed outlined in (47)-(51).

over all participants, the active inference algorithm explored on 40.48% of the trials, while the Thompson sampling algorithm and UCB explored on 3.97% and 2.20% of the trials. Note that we did not calculate the exploration behaviour of the random and optimal strategy algorithms this way, because the posterior mean does not play a role in these choice rules. Furthermore, while of theoretical interest, we did not have posterior mean estimates for the participants and could thus not calculate their percentage of exploration behaviour.

6 Discussion

6.1 Interpretation of the Algorithms

In this thesis, we set out to investigate the hypothesis that active inference models human data better than other relevant benchmark algorithms, for which we chose UCB and Thompson sampling as representatives. The choice of Thompson sampling and UCB was motivated by prior research on the two-armed bandit task (see, e.g., Gershman, 2018; Marković et al., 2021; Speekenbrink, 2019, 2022) and by the fact that both of these algorithms represent uncertainty-guided choice rules, the importance of which we take as a given in the Bayesian framework of this thesis. Hence, all three choice rules compared here are uncertainty-guided, albeit in different ways.

In Thompson sampling, the agent samples a predicted mean reward from the prior distribution and chooses the highest sampled value. This algorithm is uncertainty driven, because the variance of the estimate plays a role in the sampling process.

UCB is uncertainty-guided, because we add an uncertainty term to the mean estimate and choose the option for which this term is highest. Hence, for two options that have the same mean estimate but differing uncertainty, the algorithm chooses the one with higher uncertainty.

In this thesis' active inference implementation, an agent minimizes the sum of a risk term and an ambiguity term. The risk term comprises the Kullback-Leibler divergence between the preferences over outcomes (encoded by the Boltzmann exponential weighting function) and the predictive distribution. Analytically, in the present context, this boils down to subtracting the entropy of a t-distribution from the posterior mean estimate multiplied with the Boltzmann exploration parameter η. This Kullback-Leibler divergence has been interpreted as ensuring that those actions tend to be taken whose probabilistic outcomes match the preferred outcome (Parr et al., 2022). This also holds for the present implementation. We can see from (33) that the Kullback-Leibler divergence is closest to zero when the entropy of the t-distribution $H[T(\tilde{\lambda}(a), \tilde{\mu}(a), \tilde{\nu}(a))]$, which encodes the uncertainty associated with the predictive distribution, is equal to the Boltzmann-weighted utility term $\eta\tilde{\mu}(a)$. This can be interpreted as saying if an option has the potential of high reward, one should be willing to accept a higher degree of uncertainty in the predictive distribution (as it might be worth the risk). Of course, the opposite is also true: if an option does not have the potential of high reward, it is not worth taking high risks.

The ambiguity term encodes the expected inaccuracy that is caused by the dispersion of the distribution of outcomes (Parr et al., 2022). It is thus a measure of the mapping of states and outcomes. If this mapping is ambiguous, then we cannot learn much from choosing the respective action, making that choice less attractive. Ambiguity thus encodes the information-gain of a specific action. Hence, if the ambiguity is high, not much can be learned and this will reduce the probability of the agent choosing that option. In order to understand why, a closer look at the ambiguity term $\mathbb{E}_{Q(\mu,\lambda|a_t=a)}\left[H[\mathcal{N}(o_t|\mu, \lambda^{-1})]\right]$ is in order. Its analytical solution in the present context, $\frac{1}{2}\ln(2\pi e) + \log(\beta(a)) + \psi(\alpha(a))$, involves two expressions that depend on actions, namely $\log(\beta(a))$ and $\psi(\alpha(a))$. These two expressions also involve parameters of the gamma distribution, which models the precision λ of our learning rule. The hyperparameter β models the scale of the gamma distribution, which controls the spread - with a larger β causing a decreased height of the distribution and a further spread to the right. Thus, a lower β indicates a higher certainty in the observations as λ is expected to have less variance. This can also be reformulated as saying that the information provided in the data suggests a more tightly constrained precision. The α parameter models the shape of the gamma distribution, with higher values leading to a more peaked distribution - indicating that the data tells us more about the precision. Taken together, then, these terms describe how the data the agent observes via a particular action impacts the expected uncertainty of the normal distribution and thus the generative model (Parr et al., 2022).

In summary, the active inference choice rule presented here can be stated in plain words as: consider the trade-off between how valuable the outcome associated to a specific choice is and how uncertain it is that this outcome is obtained and furthermore consider what can be learned about the environment when taking that option. Of course, a similar choice rule could be achieved via different routes, but the point here is not only that the rule makes sense. It is important that this choice architecture is the outcome of a long list of derivations presented in this thesis that starts with belief-updating rules, which are considered plausible cognitive mechanisms by many (see section 2.2).

The parsimonious integration of the risk and ambiguity term is what sets active inference apart from the other choice rules considered here.[8] While all considered choice rules take uncertainty into account, only active inference explicitly encodes an information-gain term. UCB is based on the premise that the more variance associated with an option, the more can be learned, which should increase the probability of taking that course of action. Thompson sampling integrates the variance via its sampling method, where a high variance can increase or decrease the probability of taking an action, depending on the respective draw. Hence, these choice rules take a form of risk assessment into account, while implicitly assuming that this reveals something about the mapping between states and outcomes (ambiguity/information gain). Active inference on the other hand explicitly also considers the uncertainty in the state outcome mapping and thus the information to be gained from the respective choice (Parr et al., 2022).

6.2 Interpretation of the Results

Firstly, we want to stress that while we are using the environment of Gershman's (2018) experiment 2, a direct comparison to his results and analysis is not possible. This is because he compared the algorithms using the probit regression method, while we compared the algorithms in terms of the relative likelihood, using the AIC statistic. Using his environment nevertheless serves the purpose of standardizing a new and emerging literature on the plausibility of active inference as a cognitive mechanism.

Regarding the optimized parameters' distribution, we found it surprising that all algorithms showcased rather high memory values with active inference, UCB, Thompson sampling and the random agent having mean memory values of M = 0.50, SD = 0.22, M = 0.60, SD = 0.33, M = 0.45, SD = 0.20 and M = 0.47, SD = 0.18 respectively. This was surprising to us, because we reasoned that a good memory would store misleading information, as the mean of the reward distribution changes every 10 trials. Furthermore, in Figure 5 we can observe that at least for the active inference algorithm in tendency a lower memory value provides higher reward (except for the last step). While reward is not the optimization goal of the maximum likelihood optimization algorithm, humans perform quite well (judged by the proximity and similarity of the human data graph to the optimal performance graph in Figure 2) and it is therefore surprising that the memory parameter values that optimize performance do not optimize model fit to human data. This high memory value might partly account

[8] We consider this to be a parsimnonious formulation, because it follows from simple, generally accepted principles - as outlined in section 2.2.

for the tendency of Thompson sampling and UCB to follow trends of individual arms longer than the alternative algorithms and humans. The reason is that high memory makes these algorithms less agile to detect mean reward changes. All else being equal, this holds less for the active inference agent, because its choice is also determined by the eta parameter, which is distributed with the sample mean $M = 0.44$, $SD = 0.18$ and is thus close to the optimal performance eta value according to Figure 5. This eta value incentivises the active inference algorithm to explore quite a bit, which we can see from the rather high percentage of trials in which the active inference agent explores (40.48%) compared to Thompson sampling and UCB (3.97% and 2.20% respectively). Of course, the reason why the eta value can influence the exploration behaviour is that it determines the relative influence of two uncertainty terms - the dispersion of the predictive distribution and the information gain, or ambiguity term.

We were furthermore surprised by the fact that although active inference choices fit human choices comparably well and the curves have a similar shape (Figure 2), humans on average receive substantially more reward, as can be seen from the 95% confidence intervals not overlapping anywhere other than right at the beginning of the experiment. We believe this might have to do with the learning rule and the memory parameter. Humans are instructed that the means of the reward distributions of both arms change every 10 trials and hence they might consciously be able to not be too influenced by previous data-points, resulting in less exploration to be needed to find the optimal arm. Our exponential sliding window approach was aimed at implementing this behaviour, but was limited by the high memory values. The average memory value of active inference agents of 0.50 substantially restricts the algorithm's performance in terms of reward (as can be seen in Figure 5). This is because the algorithm will have to sample for longer in order to infer the current mean as it is influenced by data points that are irrelevant for that task.

Based on the maximum likelihood estimation the main result of this thesis is that, as hypothesized, active inference models human behaviour better than other classic uncertainty-driven choice rules. Consider again the setup of the comparison of the algorithms active inference, UCB and Thompson sampling: they all use the same Bayesian inference learning rule and they all take uncertainty into account when making a choice. However, as discussed above, while all compared algorithms involve a type of risk assessment, only active inference explicitly involves an assessment of the information gain corresponding to a specific choice. Furthermore, while the risk assessments of UCB and Thompson sampling are based on rather simple heuristics, the active inference risk term is based on variational inference - as derived in section 2.2. It is plausible that integrating uncertainty in this multifaceted way leads to more exploration behaviour, since there is constantly much information to be gained from sampling different arms, as the means are continuously changing.

Although we do not have data on the exploration behaviour of humans, we can compare the shape of their reward curve from Figure 2 to the mean of the two reward mean curves in Figure 3 and see that in tendency they do not follow local trends, but rather mirror the overall environmental reward dynamic. This, and the likelihood fit to the active inference agent indicates that humans explore quite a bit as well, as they would otherwise sometimes get stuck in local trends. This is probably the reason why the random agent has a comparably good likelihood fit: due to chance, it will explore (defined as not taking the arm with the currently highest posterior mean estimate) rather often as well.

In the context of the present dynamically changing environment, exploration thus yields benefits - both in terms of maximizing likelihood and reward. The fact that the computational setup of active inference leads to a pattern of behaviour that results in higher likelihood values can be considered as evidence that humans also either use variational inference in their risk assessment, or explicitly consider the information gain an option yields - or both. Considering the information gain seems straightforwardly plausible, since it constitutes an efficient use of a generative model that we assume the agent to have anyway and furthers efficient navigation of the environment (by reducing redundant exploration). However, the use of variational inference in planning processes is not self-evident. While it follows parsimoniously from Bayesian updating rules - as shown in section 2.2 -, there are plausible alternatives. One such alternative is the fast and frugal heuristics framework, which suggests that instead of complicated internal representations and computations, planning and decision processes can be based on rather simple rules that focus on an option being good enough rather than optimal (Conlin, 2009; Gigerenzer and Todd, 1999). One might argue that Thompson sampling and UCB can be considered as heuristic approaches and that this thesis in this sense compared variational and heuristic approaches, finding evidence in favor of the variational approaches. While this is true, we did not conduct an exhaustive comparison including all relevant heuristics and hence cannot make conclusions about the class of heuristics overall. What these two approaches have in common is that they both represent approximations to computational problems that are assumed to be too complicated for the cognitive apparatus. However, it is beyond the scope of this paper to provide a definitive answer on which mechanisms are precisely used in the human brain. The goal of this paper was rather to argue that the active inference choice rule should be considered as a plausible candidate.

A promising avenue for further research would be to combine computational modelling with imaging studies in order to find candidate neuronal correlates of active inference. Work at the intersection of computational science and neuroscience has already proven very fruitful in recent decades. Examples are the development of the reinforcement learning framework - which was inspired by the work of psychologist B.F. Skinner, who believed that we learn best when our actions are reinforced - and is now a foundational part of artificial intelligence research (Millidge, 2021; Skinner, 1958). Furthermore, we want to stress that this work has the potential to contribute to research on the alignment problem in artificial intelligence. The alignment problem is concerned with the possibility of an artificial intelligence pursuing goals that are not aligned with human goals or values (Darte and Robert, 1994). The idea is that it might be beneficial to investigate artificial intelligence algorithms that resemble the brain's computations, because if artifical intelligence algorithms and humans share a large proportion of latent code, this could lead to empathy emerging as a mis-generalization of attributing reward to another agent rather than the agent itself being in a certain state (Millidge, 2023).

7 Conclusion

In this thesis, we asked what mechanisms are involved in the human brain when faced with planning actions in an uncertain environment. Specifically, we wanted to investigate whether a theory that has received much scholarly attention in recent years, namely the free-energy principle, can better account for human behaviour than other classic reinforcement learning algorithms that deal with uncertainty. In that endeavour, we turned to a process theory of the free-energy principle – active inference – and derived as well as implemented an active inference agent, which we developed specifically for the present task design. The environment that we chose is taken from a restless bandit design from Gershman (2018). We compared this algorithm to Thompson sampling and UCB, which are commonly used algorithms in comparing restless bandits as they both deal with uncertainty, albeit in different ways. Our primary method of analysis was maximum likelihood estimation, in which we optimized each algorithm's parameters based on the likelihood fit to human performance of the identical task. Analysing the distribution of the memory parameter yielded the surprising result of rather high memory values, which impeded performance of the algorithms when solving the task using the individually optimised parameter values. Future research should investigate alternative learning rule implementations such as the Kalman filter to investigate whether model fit and performance can be optimized at the same time. We found that, in line with our hypothesis, active inference has the best likelihood fit to human data. This is evidence that the brain uses computational mechanisms that are similar to the ones involved in active inference. Specifically, this could mean that humans use variational approaches in their risk assessment or that they explicitly encode information gain terms when planning – or both. However, more research is needed to isolate the effects of these two mechanisms in order to provide definitive answers. Understanding active inference and its potential role in the cognitive machinery can contribute to our understanding of human agency and the mechanisms involved in decision processes. The applications of such research are vast and range from informing neuroscientific studies to artificial intelligence problems. The present thesis does not provide definitive answers regarding the precise role of active inference in the cognitive apparatus, but it serves as foundational research that establishes the plausibility of the hypothesis that the brain might do something similar to active inference.

References

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Bodenham, D. (2012). *Adaptive filtering and change detection for streaming data* (Doctoral dissertation).

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel'touch that eyes see. *Nature*, *391*(6669), 756–756.

Bruineberg, J., Dołega, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new markov blankets. *Behavioral and Brain Sciences*, *45*, e183. https://doi.org/10.1017/S0140525X21002351

Cesa-Bianchi, N., Gentile, C., Lugosi, G., & Neu, G. (2017). Boltzmann exploration done right. *Advances in neural information processing systems*, *30*.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204. https://doi.org/10.1017/S0140525X12000477

Colombo, M., Elkin, L., & Hartmann, S. (2021). Being realist about bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*.

Conlin, J. A. (2009). Getting around: Making fast and frugal navigation decisions. *Progress in brain research*, *174*, 109–117.

Darte, A., & Robert, Y. (1994). On the alignment problem. *Parallel Processing Letters*, *4*(03), 259–270.

Ehrsson, H. H. (2007). The experimental induction of out-of-body experiences. *Science*, *317*(5841), 1048–1048.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in cognitive sciences*, *13*(7), 293–301.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural computation*, *29*(1), 1–49.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*, 187–214. https://doi.org/10.1080/17588928.2015.1020053

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gigerenzer, G., & Todd, P. M. Fast and frugal heuristics: The adaptive toolbox. In: In *Simple heuristics that make us smart*. Oxford University Press, 1999, pp. 3–34.

Helmholtz, H. V. (1860). *Handbuch der physiologischen optik [english translation]* (1962nd ed.). Dover.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Hohwy, J. (2020). New directions in predictive processing. *Mind Language*, *35*, 209–223. https://doi.org/10.1111/MILA.12281

Kant, I. (1999). *Critique of pure reason (the cambridge edition of the works of immanuel kant)*. Cambridge University Press.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of The royal society interface*, *15*(138), 20170792.

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean journal of anesthesiology*, *70*(2), 144–156.

Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, *317*(5841), 1096–1099.

Marković, D., Stojić, H., Schwöbel, S., & Kiebel, S. J. (2021). An empirical evaluation of active inference in multi-armed bandits. *Neural Networks*, *144*, 229–246.

Millidge, B. (2021). Applications of the free energy principle to machine learning and neuroscience. *arXiv preprint arXiv:2107.00140*.

Millidge, B. (2023). *Empathy as a natural consequence of learnt reward models*. Retrieved August 2, 2023, from https://www.lesswrong.com/posts/zaER5ziEprE7aNm6u/empathy-as-a-natural-consequence-of-learnt-reward-models

Olivé, I., & Berthoz, A. (2012). Combined induction of rubber-hand illusion and out-of-body experiences. *Frontiers in Psychology*, *3*, 128.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Powell, M. J. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, *7*(2), 155–162.

Seth, A. (2021). *Being you*. Faber Faber Ltd.

Skinner, B. F. (1958). Reinforcement today. *American Psychologist*, *13*(3), 94.

Soch, J., of Statistical Proofs, T. B., Faulkenberry, T. J., Petrykowski, K., Allefeld, C., & McInerney, C. D. (2022, January). *StatProofBook/StatProofBook.github.io: StatProofBook 2021* (Version 2021). Zenodo. https://doi.org/10.5281/zenodo.5820411

Speekenbrink, M. (2019, January). Modeling reinforcement learning (part ii): Maximum likelihood estimation. https://doi.org/10.17605/OSF.IO/SJHUY

Speekenbrink, M. (2022). Chasing unknown bandits: Uncertainty guidance in learning and decision making. *Current Directions in Psychological Science*, *31*(5), 419–427.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3-4), 285–294.

Tschantz, A., Millidge, B., Seth, A. K., & Buckley, C. L. (2020). Reinforcement learning through active inference. http://arxiv.org/abs/2002.12636

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review*, *11*, 192–196.

Wakayama, S., & Ahmed, N. Active inference for autonomous decision-making with contextual multi-armed bandits. In: In *2023 ieee international conference on robotics and automation (icra)*. IEEE. 2023, 7916–7922.

Appendix A

To reduce clutter, we will suppress conditioning on $a_t = a$ and the fact that the functions are parameterized by a. Starting with the given expressions:

$$\begin{aligned}
Q(\mu, \lambda) &= \text{Normal-Gamma}(\mu, \lambda | \mu_0, \lambda_0, \alpha, \beta) \\
&\propto \exp\left(-\frac{\lambda\lambda_0}{2}(\mu - \mu_0)^2\right) \lambda^{\alpha - \frac{1}{2}} \exp(-\beta\lambda) \\
Q(o) &= \int Q(o|\mu, \lambda) Q(\mu, \lambda)\, d\mu\, d\lambda \\
&\propto \int \sqrt{\lambda} \exp\left(-\frac{\lambda(o-\mu)^2}{2}\right) \exp\left(-\frac{\lambda\lambda_0}{2}(\mu - \mu_0)^2\right) \lambda^{\alpha - \frac{1}{2}} \exp(-\beta\lambda)\, d\lambda\, d\mu \\
&= \int \lambda^{\alpha} \exp\left(-\frac{\lambda}{2}\left[2\beta + (o-\mu)^2 + \lambda_0(\mu - \mu_0)^2\right]\right) d\lambda\, d\mu
\end{aligned}$$

Simplifying the quadratic expression in the exponential:

$$\begin{aligned}
(o-\mu)^2 + \lambda_0(\mu - \mu_0)^2 &= (1 + \lambda_0)\mu^2 - 2\mu(o + \lambda_0\mu_0) + o^2 + \lambda_0\mu_0^2 \\
&= (1 + \lambda_0)\left(\mu - \frac{o + \lambda_0\mu_0}{1 + \lambda_0}\right)^2 + u \\
\text{where} \quad u &= \frac{1}{2}\left(o^2 + \lambda_0\mu_0^2 - \frac{(o + \lambda_0\mu_0)^2}{1 + \lambda_0}\right).
\end{aligned}$$

Integrating the exponential terms:

$$\begin{aligned}
&\int \exp\left(-\frac{\lambda}{2}\left[(o-\mu)^2 + \lambda_0(\mu - \mu_0)^2\right]\right) d\mu \\
&\propto \int \exp\left(-\lambda(1 + \lambda_0)\left(\mu - \frac{o + \lambda_0\mu_0}{1 + \lambda_0}\right)^2 - \lambda u\right) d\mu \\
&\propto \exp(-\lambda u)\sqrt{\lambda}
\end{aligned}$$

Continue the integration:

$$\begin{aligned}Q(o) &\propto \int \lambda^{\alpha+\frac{1}{2}} \exp(-\lambda u - \lambda\beta)\, d\lambda \\ &= \int \left(\frac{x}{u+\beta}\right)^{\alpha+\frac{1}{2}} \exp(-x)\frac{1}{u+\beta}\, dx \\ &\propto \frac{1}{(u+\beta)^{\alpha+\frac{3}{2}}}\end{aligned}$$

expressing u in a simplified form:

$$u = \frac{\lambda_0}{2}\frac{(o-\mu_0)^2}{\lambda_0+1}$$

From the simplification of u and the resulting expression for $Q(o)$, we find that:

$$Q(o) \propto \left(1 + \frac{\lambda_0}{2\beta(1+\lambda_0)}(o-\mu_0)^2\right)^{-\frac{2(\alpha+1)+1}{2}}$$

After integration and simplification, the expression for $Q(o)$ involves a term $\frac{\lambda_0}{2\beta(1+\lambda_0)}(o-\mu_0)^2$. Deconstructing this term, it becomes apparent that $\frac{\lambda_0}{2\beta}$ shares an analogy with the shape parameter of the Gamma distribution, while $\frac{(o-\mu_0)^2}{1+\lambda_0}$ closely resembles a scaled chi-squared variable with 1 degree of freedom and a scaling factor $\frac{1}{1+\lambda_0}$.

By recognizing this resemblance and considering the form of a non-central t-distribution density function:

$$f(x;\nu,\mu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}$$

We deduce the following t-distribution parameters:

$$\begin{aligned}\text{Degrees of Freedom: } &\tilde{\nu} = 2(\alpha+1) \\ \text{Location Parameter: } &\tilde{\mu} = \mu_0 \\ \text{Scale Parameter: } &\tilde{\lambda} = \frac{\beta(1+\lambda_0)}{\lambda_0(1+\alpha)}\end{aligned}$$

Hence, the derived posterior expression $Q(o)$ possesses the characteristics of a non-central t-distribution, and the identified parameters reflect its degrees of freedom, location, and scale. This derivation, inspired by Bishop and Nasrabadi (2006), thus establishes a link between the variational Q distribution and the t-distribution parameters within the context of Bayesian analysis.

Appendix B

Participant	AI Likelihood	AI AIC	ucb Likelihood	ucb AIC	thompson Likelihood	thompson AIC	random Likelihood	random AIC
1	136.460309183725	239.933563822132	265.56957232192	162.757611827948	276.92061836745	483.867127644264	533.139144643839	329.515223655895
2	138.127620036926	750.524372900253	687.619136227462	162.012776444498	280.255240073852	1505.04874580051	1377.23827245492	328.025552888996
3	136.600539582696	227.176787230833	197.071535366961	165.580367394516	277.201079165393	458.353574461666	396.143070733923	335.160734789031
4	130.941498828616	467.97128904883	528.102078862713	162.517590082585	265.882997657232	939.942578097661	1058.20415772543	329.035180165169
5	136.319312945222	404.065800580485	272.708487475192	158.745708445157	276.638625890444	812.13160116097	547.416974950385	321.491416890315
6	139.342959621555	759.037462830834	900.425231563479	163.671646494512	282.68591924311	1522.07492566167	1802.85046312696	331.343292989024
7	125.963846951323	357.027523103106	441.704404008279	159.895137952008	255.927693902645	718.055046206213	885.408808016558	323.790275904015
8	135.55804255481	277.894356247636	292.142137246018	158.556869565381	275.11608510962	559.788712495272	586.284274492035	321.113739130761
9	132.514816613308	607.666595643713	453.507617898601	167.822450428532	269.029633226616	1219.33319128743	909.015235797202	339.644900857063
10	122.599680984504	229.735117397087	300.547897653834	162.39360402492	249.199361969007	463.470234794175	603.095795307668	328.787208049839
11	140.135071375078	870.333839966476	486.032897038283	151.732175676025	284.270142750156	1744.66767993295	974.065794076567	307.46435135205
12	135.509797447065	716.150979478459	607.160175650457	161.95594695431	275.01959489413	1436.30195895692	1216.32035130091	327.911893908619
13	134.422094365021	434.868115570772	172.786287001375	159.848605121414	272.844188730042	873.736231141544	347.57257400275	323.697210242828
14	126.271523003935	300.99144276672	381.165904926295	162.288710948713	256.54304600787	605.982885533441	764.33180985259	328.577421897426
15	125.525096127518	426.23109240257	371.409516613402	162.485432003107	255.050192255036	856.46218480514	744.819033226805	328.970864006215
16	141.96784785685	961.905130672031	661.10797270403	162.753817559315	287.935695713701	1927.81026134406	1324.21594540806	329.507635118629
17	135.267097604315	274.412954038048	355.472196304609	154.666146269641	274.53419520863	552.825908076096	712.944392609218	313.332292539281
18	134.334231483262	457.167674993691	318.177092761189	167.938883960617	272.668462966524	918.335349987382	638.354185522378	339.877767921235
19	139.78527572498	281.757055416929	299.410605918686	166.038338043495	283.57055144996	567.514110833858	600.821211837373	336.076676086989
20	132.041980395806	283.448079263709	238.715800508141	157.702136564474	268.083960791612	570.896158527417	479.431601016282	319.404273128948
21	127.603414469728	296.885541131597	254.775924577768	163.290980533453	259.206828939456	597.771082263193	511.551849155536	330.581961066907
22	133.027335684839	329.147524584719	349.14219948103	159.264944311507	270.054671369678	662.295049169438	700.28439896206	322.529888623015
23	141.369682648311	247.677858815072	239.275913259864	162.66299606301	286.739365296621	499.355717630144	480.551826519728	329.325992126019
24	130.502098197627	348.54920512179	342.741619271753	153.114877022616	265.004196395254	701.09841024358	687.483238543507	310.229754045232
25	135.113806379157	295.611888983143	321.942419105838	156.039553730711	274.227612758314	595.223777966285	645.884838211677	316.079107461422
26	133.577551453303	203.125601511303	292.092847942468	166.381109351148	271.155102906605	410.251203022605	586.185695884936	336.762218702295
27	135.99448055767	441.104227634813	346.5593799816	156.720347759068	275.988961115339	886.208455269626	695.118759963199	317.440695518137
28	140.568757638456	291.91952992193	297.924349199901	165.644137093695	285.137515276913	587.83905984386	597.848698399803	335.28827418739
29	134.792942019045	569.275959857665	272.150858315693	168.053375002736	273.58588403809	1142.55191971533	546.301716631385	340.106750005471
30	134.378750237031	683.58623207092	584.419340295163	149.60945085191	272.757500474061	1371.17246414184	1170.83868059033	303.218901703819
31	141.193979141286	685.528846849856	426.037280606332	169.635858685692	286.387958282573	1375.05769369971	854.074561212664	343.271717371385
32	133.271426550258	690.622801157152	931.176787023051	160.012128929461	270.542853100516	1385.2456023143	1864.3535740461	324.024257858922
33	137.064866712601	180.576822669069	211.008461589333	158.424408205211	278.129733425202	365.153645338137	424.016923178666	320.848816410423
34	133.481411389209	440.342381408084	418.323650041207	166.197803404081	270.962822778417	884.684762816168	838.647300082415	336.395606808162
35	132.795164338584	877.091399230271	416.556101199519	153.689783223113	269.590328677167	1758.18279846054	835.112202399038	311.379566446225
36	142.008287177746	265.353321142496	213.904224555623	153.73747194442	288.016574355492	534.706642284992	429.808449111247	311.47494388884
37	135.394004116432	625.235151015042	229.100662770803	167.488508009257	274.788008232864	1254.47030203008	460.201325541606	338.977016018514
38	131.314277660671	323.602462380069	304.020901797871	160.599485630356	266.628555321342	651.204924760137	610.041803595741	325.198971260713
39	132.795692567339	599.626942584966	581.164362541422	155.673519950527	269.591385134677	1203.25388516993	1164.32872508284	315.347039901054
40	137.437807300965	475.984156647652	550.680906023493	165.219319905145	278.87561460193	955.968313295304	1103.36181204699	334.43863981029
41	134.276199290859	272.51703307668	303.820911068152	163.837064849121	272.552398581719	549.03406615336	609.641822136303	331.674129698241
42	138.949580510466	1135.17833801	820.758337465493	162.590637431071	281.899161020932	2274.35667602	1643.51667493099	329.181274862142
43	135.043962562554	727.484926335087	736.067961583567	162.565760387617	274.087925125107	1458.96985267017	1474.13592316713	329.131520775234
44	134.133331356905	579.999620380257	544.267452044295	164.830759810819	272.26666271381	1163.99924076051	1090.53490408859	333.661519621637

Table 1: Maximum Likelihood and AIC Statistics per Participant per algorithm