A deep active inference model of the rubber-hand illusion

Bachelor's Thesis in Artificial Intelligence by

Thomas Rood[1]
s1005156.

Supervised by:
Pablo Lanillos[1,2]

Second reader:
Marcel van Gerven[1,2]

[1]Department of Artificial Intelligence, Radboud University
[2]Donders Institute for Brain Cognition and Behaviour, Radboud University

July 12, 2020

Abstract

How human perception and action deal with uncertainty has been extensively studied using the rubber-hand illusion (RHI), which introduces a sensorimotor conflict by means of a rubber hand, causing a shift in the perceived location of the real hand. A recent study showed that in addition to the drift in perception, the RHI also evokes an involuntary force at the hand in the direction of rubber hand. Seemingly congruent to this finding is the strong interdependence between perception and action found in active inference. In this work we propose a deep active inference agent which deals with high-dimensional visual inputs in a virtual environment, that is able to account for the action produced by humans under the RHI. In addition, the agent was subjected to reaching tasks, validating the proper function of the agent's perception and action.

I. Introduction

The complex mechanisms underlying perception and action that allow seamless interaction with the environment are largely occluded from our consciousness, yet crucial to our existence. To interact with the environment in a meaningful way, the brain must integrate noisy sensory information from multiple modalities into a coherent world model, from which to generate and continuously update an appropriate action [1]. How the brain realises this is an important longstanding question in cognitive science and artificial intelligence, resulting in a variety of models.

One such model, the Bayesian inference model, states that perception and action (and their interaction) can be approached using conditional probability density functions [2]. The neurophysiological and behavioural predictions this model makes have seen a wide range of support from empirical data (e.g. [1], [3], [4], [5]), as well as proofing useful in artificial implementations such as computer vision [6].

The free-energy principle is a recent framework build upon this theory of Bayesian inference. Inspired by Helmholtz's view on perception [7], Karl Friston put forward this principle as a fundamental and uniform theory of the brain's functioning [8], [9], [10]. The idea underlying the free-energy principle is that biological systems actively try to minimise the disparity between their prediction of sensory input and the actual sensory input [9]. This disparity, known as 'surprise', depends on the prediction generated by the system's internal model and drives the adaptation of brain states (perception) and the internal model (learning) [11].

The free-energy principle differentiates itself from other theories by postulating that the minimisation of surprise not only occurs through perception and learning, but also through action in the environment [10]. It describes that, because action changes the state of the environment, it can be used to indirectly change future sensory input in an attempt to minimise the prediction error. This framework that directly couples the optimisation of perception, action and learning has recently gained attention from the scientific community, yielding promising results in image quality metrics [12], [13], [14] and robotic perception and action [15], [16], [17].

Although the free-energy principle is presented as a unified theory of the brain [10] and shows promising results in artificial applications, how can its capacity to accurately model human behaviour be assessed?

A. Rubber-hand illusion (RHI) and computational models

This problem can be approached by comparing the models' behaviour to that of humans under a well-known experimental paradigm in cognitive science: the rubber-hand illusion (RHI) [18]. This paradigm is used to examine the properties of body self-perception by presenting a rubber hand and providing sensory evidence that this hand is ones own. This evidence is provided in the form of synchronous visuo-tactile stimulation, which in the original experiment took the form of a paintbrush stroking the real and rubber hand synchronously [18]. The paradigm shows that within under a minute of stimulation [19], the perceived location of the real hand drastically shifts towards the location of the rubber-arm (proprioceptive drift), and vise versa (visual drift) [20].

In [21], a Bayesian causal inference model was proposed that used visual, tactile and proprioceptive sensory input and their respective precision to estimate the hand position when subjected to the RHI. They found that the model produced results similar to the results produced by human subjects under the RHI. Furthermore, their model predicted that the RHI could occur in the absence of tactile stimulation, which they confirmed in a subsequent experiment on humans. These results suggest that Bayesian causal inference is an important part of human body self-perception [21].

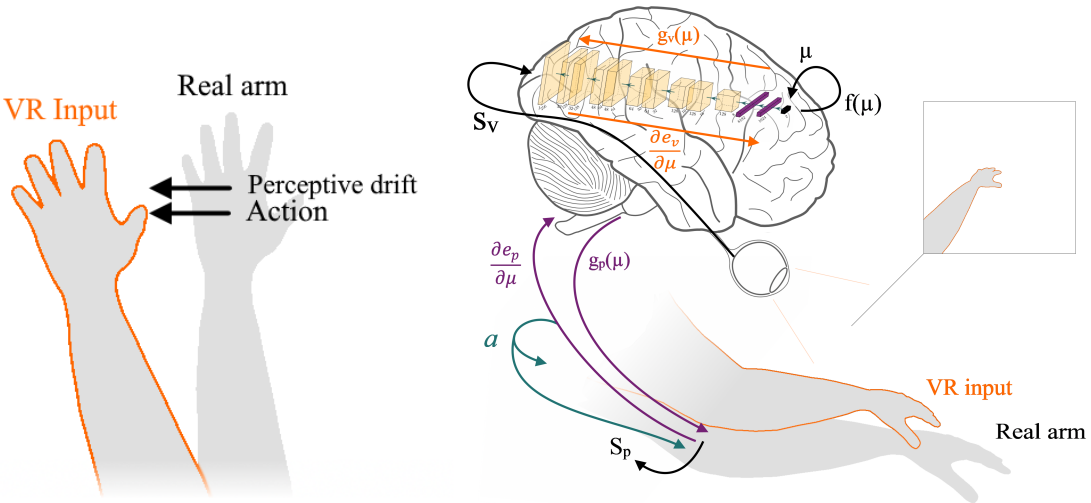
Similarly oriented models have been proposed for the free-energy principle. For example, the literature has seen a robotic implementation of the free-energy principle [22], which was subjected to the RHI. The perceptual drift was compared to that of humans and found to be similar, supporting the free-energy principle as a viable model of body self-perception.

These models all make use of multimodal perception to estimate the current body state, showing how the RHI has an effect on passive perception. However, a recent experiment showed how humans unconsciously produce a small involuntary force at the hand in the direction of the rubber hand [23]. This finding adds a new dimension to the RHI, forming an additional requirement on synthetic models that attempt to predict human behaviour under the RHI.

B. Contribution

To our knowledge, there is currently no model implemented in the literature that is able to account for these findings. We hypothesise that the aforementioned strong interdependence between perception and action found in the free-energy principle fits these new findings, and is therefore a viable candidate for a synthetic model of the RHI. This means that we expect the model to produce action in the direction of a rubber hand during the RHI, as well as observing a shift in the perceived location of the real hand – See Fig. 1a.

In this work[1], a new synthetic model for the free-energy principle operating in a 3D virtual environment is proposed based on the algorithm proposed by [16] and the findings of [23]. Our aim was to create a model to account for the forces generated during the RHI [23], as well as providing a new and scalable testing ground for further research on the free-energy principle[2]. The proposed model will therefore include the implementation of visual perception using raw imagery originating from the 3D environment. The model will be subjected to the RHI and its output will be compared to the findings of [23]. Our expectation is that the model will produce similar action patterns in the direction of the rubber hand. In addition, reaching tasks will be performed to assess the validity of the actions produced by the model.



(a) (b)

Fig. 1: (a) Observed effects in the RHI experiment. (b) Deep active inference model for the virtual rubber-hand illusion. The brain variables μ that represent the body state are inferred through proprioceptive s_p and visual s_v prediction errors and their own dynamics $f(\mu)$. During the VR immersion, the agent only sees the VR arm.

II. Methods

A. Deep Active Inference Model

Under the free-energy principle, perception and action are driven by a minimisation of prediction error [24]. Specifically, the free-energy principle describes a dynamic internal state which is connected to the world through sensory input. This means that the internal state can only access the body state (and environment) through the senses. The sensory input combined with the internal state are used to approximate the real body state, under the minimisation of variational free energy.

[1]This work was submitted to the 1st International Workshop on Active Inference.

[2]Full environment and agent code is publicly available at https://github.com/thomasroodnl/active-inference-rhi

The proposed agent made use of two sensory modalities. The visual input s_v is described by a pixel matrix (image) and the proprioceptive information s_p represents the joint angles of the arm – See Fig. 1b. The following paragraphs show the variational approximation of the body state distribution μ and the action a as defined by the free-energy principle under the mean-field and Laplace approximations. Note that this is a specific instance of the full free-energy principle formulation [10] tailored to the RHI experimental setup. For the full derivations of these formulas, see [25], [26], [16].

The body state belief μ and action a are determined through a minimisation of the variational free-energy on the sensory input and previous body state belief. We compute this minimisation through gradient descent This means that μ and a are updated every iteration with time-step differences $\dot{\mu}$ and $\dot{a}$ shown in Equation (3) and (4).

$$\mu = \underset{\mu}{argmin}\ F(s, \mu) \tag{1}$$

$$a = \underset{a}{argmin}\ F(s, \mu) \tag{2}$$

$$\mu_t = \mu_{t-1} + \dot{\mu} \cdot \Delta t \tag{3}$$

$$a_t = a_{t-1} + \dot{a} \cdot \Delta t \tag{4}$$

To minimise the variational free-energy at every iteration, the terms $\dot{\mu}$ and $\dot{a}$ are set to the the negative partial derivative of the free energy with respect to μ ($-\frac{\partial F}{\partial \mu}$) and a ($-\frac{\partial F}{\partial a}$).

1) Perception inference: The free-energy optimisation with respect to μ consists three components, being the optimisation with respect to the visual sensory input s_v, the proprioceptive sensory input s_p and a possible visual goal $s_{v_{goal}}$. This term was added to the free-energy formulation to be able to enforce particular actions to be performed in a model that otherwise purely searches for an equilibrium state [16]. Equation (5) shows how $\dot{\mu}$ can be rewritten as the minimisation of the free energy with respect to the three components mentioned before.

$$\dot{\mu} = -\frac{\partial F}{\partial \mu} = (-\frac{\partial F_v}{\partial \mu}) + (-\frac{\partial F_p}{\partial \mu}) + (-\frac{\partial F_{v_{goal}}}{\partial \mu}) \tag{5}$$

The visual component of the free-energy gradient with respect to μ was computed as the error between the visual sensation s_v and visual prediction $g(\mu)$, multiplied by the partial derivative of $g(\mu)$ with respect to μ to convert the error from visual space to body state space. When $g(\mu)$ takes the form of neural network, this partial derivative can easily be obtained by performing a backward pass through the network [16]. The intensity of this gradient is primarily modulated by Σ_μ^{-1}, which denotes the visual precision. To be able to control the strength of the visual component based on the strength of the RHI, gain parameter γ was added to the equation. This allowed us to separate the dynamic change in illusion strength from the static visual precision. Section II-B describes how the value of γ depends on the visuo-tactile stimulation in the context of the RHI.

$$-\frac{\partial F_v}{\partial \mu} = \frac{\partial g(\mu)}{\partial \mu} \cdot \frac{\partial F_v}{\partial g(\mu)} = \frac{\partial g(\mu)}{\partial \mu} \cdot \gamma \Sigma_v^{-1}(s_v - g(\mu)) \tag{6}$$

In similar fashion to the visual component, the visual goal part of the free-energy gradient with respect to μ was computed as the error between the visual sensation of the goal $s_{v_{goal}}$ and visual prediction $g(\mu)$, multiplied by the partial derivative of $g(\mu)$ with respect to μ to convert the error from visual space to body state space. The intensity of the gradient is modulated by scalars β and Σ_μ, which respectively denote the goal strength and the variance of μ.

$$-\frac{\partial F_{v_{goal}}}{\partial \mu} = \frac{\partial g(\mu)}{\partial \mu} \cdot \frac{\partial F_{v_{goal}}}{\partial g(\mu)} = \frac{\partial g(\mu)}{\partial \mu} \cdot \beta \Sigma_\mu^{-1}(s_{v_{goal}} - g(\mu)) \tag{7}$$

The proprioceptive part of the free-energy gradient with respect to μ was defined in a similar way as the visual part. However, as opposed to the visual part, proprioception was modelled to reside in the same domain as the body state belief μ, namely the joint angle space. This means that proprioceptive prediction boils down to belief μ. Subsequently, the proprioceptive prediction error is computed as by subtracting μ from proprioceptive input s_p, and no conversion to body state space is necessary. The intensity of the gradient is modulated by Σ_p^{-1}, denoting the proprioceptive precision.

$$-\frac{\partial F_p}{\partial \mu} = \Sigma_p^{-1}(s_p - \mu) \tag{8}$$

2) Active inference: The problem of action is formulated similarly to that of body perception, namely as gradient optimisation on the partial derivative of the free energy with respect to the action. Just as with body perception, action is updated in a way that minimises the resulting variational free energy.

In contrast to the the differential equation that drives perception, the time-step difference of action $\dot{a}$ was modelled to only of consist an optimisation of the proprioceptive free energy with respect to the action. This is coherent with the view of action as a way to cancel proprioceptive prediction error akin to classical body reflex arcs [27].

$$\dot{a} = -\frac{\partial F}{\partial a} = -\frac{\partial F_p}{\partial a} \tag{9}$$

$$-\frac{\partial F_p}{\partial a} = -\Sigma_p^{-1}(s_p - \mu) \tag{10}$$

B. Visuo-tactile synchrony

To be able to simulate the rubber-hand illusion using the discussed free-energy model, the model needed to be extended to incorporate the influence of visuo-tactile stimulation on the perceived causality of the sensory input, similar to the model used in [15]. To simplify this problem, the environment was configured to provide the time points at which a visual stimulation event and the corresponding tactile stimulation took place, which were denoted as t_v and t_t respectively. We distinguished between two explanations of causality using the variable C, approaching it as a Bayesian causal inference problem [21]. It is either the case that the observed (rubber) hand produced both the visual and the tactile event (the rubber hand is perceived as our own), denoted by $C = 1$, or the observed hand produced the visual event and our real hand produced the tactile event (the rubber hand not perceived as our own), denoted by $C = 2$. The strength of the illusion embodied by γ was formulated as the probability of $C = 1$ given time points t_v and t_t. Bayes' rule was applied to this probability in order to obtain Equation (11).

$$p(C = 1 \mid t_v, t_t) = \frac{p(t_v, t_t \mid C = 1) \cdot p(C = 1)}{p(t_v, t_t \mid C = 1) \cdot p(C = 1) + p(t_v, t_t \mid C = 2) \cdot p(C = 2)} \tag{11}$$

This equation describes how the probability of $C = 1$ is updated based on a visuo-tactile event and the previous or prior probability $p(C = 1)$, and was used in the model as the temporal updating step. The likelihood of the visuo-tactile event given $C = 1$ $p(t_v, t_t \mid C = 1)$ was defined as a Gaussian probability density function over the error between t_v and t_t with a mean of 0. The idea behind this was that when the error is zero, the likelihood should be maximal and when the error increases, the likelihood decreases.

$$p(t_v, t_t \mid C = 1) = \mathcal{N}(t_v - t_t, 0, \sigma^2) \tag{12}$$

$$p(t_v, t_t \mid C = 2) = \mathcal{U}(t_v - t_t, \Delta t_{min}, \Delta t_{max}) \tag{13}$$

The likelihood of the visuo-tactile event given $C = 2$ was defined as a uniform distribution over the error between t_v and t_t. Because $C = 2$ denotes the hypothesis that the visual and tactile events are produced by different arms, these events are independent. This means that $p(t_v, t_t \mid C = 2) = p(t_v \mid C = 2) \cdot p(t_t \mid C = 2)$. However, the probability of a visual or tactile event happening at a certain absolute time point depends on the implementation of the stimulation (e.g. stimulation at regular intervals or randomly) and is not of main interest to the RHI paradigm. Rather, the relative difference between the events matters, which is why the likelihood given $C = 2$ was modelled as a uniform probability density function of the error between t_v and t_t. The uniform distribution was limited by the minimally and maximally possible time differences Δt_{min} and Δt_{max}, which are defined by the experimental setup.

This causality computation was integrated in the model by updating the strength of the visual component γ with the value of $p(C = 1 \mid t_v, t_t)$ every time a visuo-tactile event occurs. Equation (11) was computed by simply filling in the value of γ for the current timestep as the prior $p(C = 1)$. By observing that $C = 1$ and $C = 2$ are the only causal explanations in C and are each other's opposite, the prior $p(C = 2)$ can be defined as $1 - p(C = 1)$. Thus, $p(C = 2)$ can be computed by taking $1 - \gamma_t$.

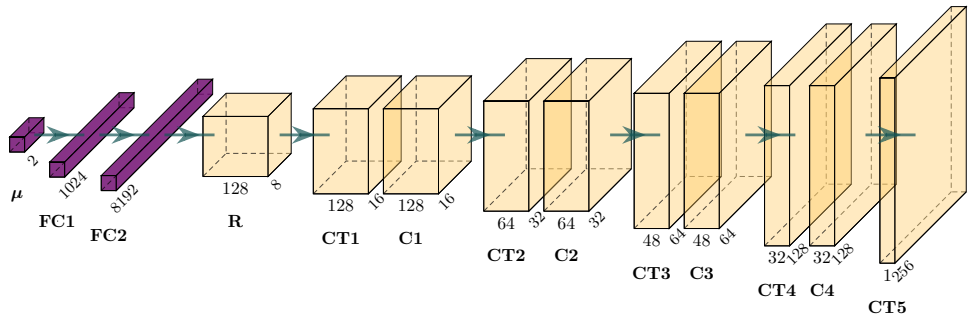
$$\gamma_{t+1} = \begin{cases} \frac{p(t_v,t_t|C=1)\cdot\gamma_t}{p(t_v,t_t|C=1)\cdot\gamma_t + p(t_v,t_t|C=2)\cdot(1-\gamma_t)}, & \text{if visuo-tactile event} \\ \gamma_t \cdot exp(-\frac{(t-t_t)^2}{\Delta_t^{-1}} \cdot r_{decay}), & \text{otherwise} \end{cases} \tag{14}$$

With the Bayesian update rule in place, the value of γ changes each time a visuo-tactile event occurs. However, without events occurring, the value of γ would simply remain stationary solely relying on this formula. This would imply that a complete lack of evidence for long time spans would have no effect on the strength of the illusion. Although the effect of the absence of evidence over time on the strength of the illusion had, to our knowledge, not been systematically studied yet, we predicted the strength of the illusion to decay over time. Our reasoning is as follows: If the illusion is strong, the system is in a constant high-error state because of the discrepancy between the proprioceptive and the visual input. This discrepancy can be justified for in the presence of clear evidence that forces the probability of $C = 1$ to be high. However, in absence of such evidence, there is no continuous justification for the high error, and as a system that minimises the free-energy, the system should lower its estimated probability of $C = 1$ to minimise error. Thus, we modelled the behaviour of γ when no event is occurring as a decay based of which the strength is based on the amount of time that has passed since the last visuo-tactile event. However, because a visuo-tactile event consists of two separate time points, we only use the time point of the last tactile event.

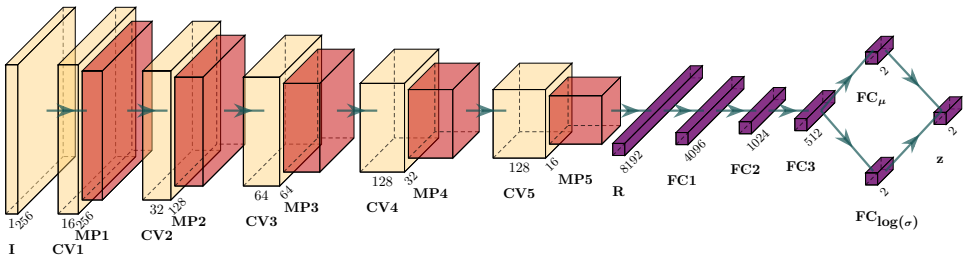
Finally, we note that the prior value of γ_t cannot be initialised to zero as this would prevent the model from ever attaining a non-zero illusion strength. Furthermore, there is evidence that humans can experience the illusion without any stimulation [21], meaning that γ must start at a non-zero value.

C. Visual forward models

To be able to use visual sensory information to update the belief and action, a visual forward model was needed to create visual predictions from which to compute the prediction error. Two different neural networks solutions for this visual model were considered, a (deep) convolutional decoder and a variational autoencoder (VAE).



(a) Convolutional decoder network architecture



(b) Variational autoencoder architecture (encoder only)

Fig. 2: Visual forward model architectures. FC: Fully-connected (dense) layer, R: Reshape operation, C: 2D Convolutional layer, CT: 2D Convolutional Transposed layer.

1) Convolutional decoder: The convolutional decoder was designed in similar fashion to the decoder used in [16] (in turn based on architectures used in [28]), which showed promising results in generating forward visual predictions and allowed for the intuitive computation of the complex relation between the visual space and the joint space $\frac{\partial g(\mu)}{\partial \mu}$ by means of the backward pass through the network. The decoder employed two fully connected layers to generate a large hidden representation vector from μ, which was reshaped into a 3-dimensional tensor. This tensor was transformed into a grayscale image by alternately applying upconvolutional and convolutional operations, reducing the amount of kernels and increasing the resolution (see Fig. 2a), while reducing visual artifacts respectively [28]. The ReLU activation function was used for all layers except the last one, which employed a sigmoid activation function to squeeze the output in the [0, 1] range complying with the grayscale imagery. Our implementation of the decoder was designed with the experimental setup in mind, noting that the different dimensions of our human model compared to the NAO robot model used in [16] would require a different approach. The human model being multiple magnitudes larger than the NAO robot meant that the arms were relatively further away from the camera, resulting in sparse camera images. To retain proper rendering of the arms in the visual prediction, a high resolution of 256x256 pixels was chosen as the visual prediction output. To achieve this, the main decoder proposed in [28] was adjusted by adding an additional set of upconvolutional and convolutional layers and changing the number of kernels per layer to accommodate this new layer (See Fig. 2a).

2) Variational autoencoder: The variational autoencoder was designed using the same decoding structure as the convolutional decoder to allow for the comparison of their respective performance. This meant that after training, these models operated equivalently, yet their different training architecture could prove important to model performance. The variational autoencoder differs from the convolutional decoder by using an encoder structure during training that encodes a latent representation of the input image I in the form of a mean and variance vector (see Fig. 2b). These vectors are used to constitute a multivariate Gaussian distribution from which to sample the estimated belief z, which is fed into the aforementioned decoder structure to produce output image $\hat{I}$.

3) Training: To train the two visual models, 8000 random arm positions were generated by uniformly sampling random angles for both the shoulder and elbow within a range of [-50, 50] degrees relative to the center condition (see section II-D3). The true arm joint angles q and the corresponding visual perception s_v of these generated positions were collected and made up the training and validation set (10% of total entries used for validation).

The convolutional decoder was trained by providing the network with true arm positions from the training dataset and letting the network produce visual outputs. The loss was computed as the mean squared error (MSE) between the visual outputs and the visual ground truths (corresponding to the true arm positions provided to the network as input), which was used to perform gradient descent on the weights of the network.

The variational autoencoder was trained by providing the network with the visual images from the training dataset and letting the network produce visual outputs. The loss was computed as the sum of two terms. The first term was again the mean squared error (MSE) loss between the visual ground truth (in this case also the input to the model I) and the visual prediction ($\hat{I}$). The second term was a regularisation term $D_{KL}(q(z \mid I) \mid\mid p(z))$ being Kullback-Leibler divergence between the distribution $q(z \mid I)$ embodied by the encoder's mean and variance vectors and desired distribution $p(z)$ [29]. By setting $p(z)$ to a Gaussian distribution with as mean the true arm position and as variance some small value (i.e. $\mathcal{N}(x, 0.001)$), the latent space was enforced to resemble the joint angle space. This way, the decoder part of the network could be used to make visual prediction from joint angle beliefs during agent operation.

Both visual models were trained using the ADAM optimiser with an initial learning rate 0.001. This learning rate was decayed every 20 epochs with a rate of 0.95. Mean squared error was used as a loss criterion, and a minibatch size of 200 was used. The visual models were both trained for 800 epochs.

D. Experimental setup

1) Free-energy model: The free-energy model discussed in section II-A was implemented in Python building upon the PixelAI algorithm from [16]. The most notable extensions of this algorithm were the inclusion of the proprioceptive action component, the illusion strength algorithm and the integration with the Unity engine simulation environment (see next section).

2) Environment: The experiment took place in a virtual environment modelled in the Unity engine [30]. For the agent, the same configurable arm model as used in [23] was used. To simplify the free-energy optimisation and visual decoder training, the free-energy agent was given velocity control over the joints of the left arm, which was granted two degrees of freedom, being shoulder adduction and abduction (moving towards and away from the body)

and elbow flexion and extension (moving the lower arm towards and away from the upper arm). The body state space was subsequently defined as the angles of these degrees of freedom with respect to the initial position.

The environment provided proprioceptive sensory input on the aforementioned joint angles to the free-energy model in degrees relative to the initial joint angle position, as defined by the RHI paradigm (see next section). Visual sensory input to the model originated from a camera located between the left and right eye positions, producing 256x256 pixel grayscale images. This camera was configured to not be able to observe the real body, but instead only see the presented 'rubber' body. A solid grey background was used to avoid visual interference and decoder artefacts originating from properties other than the presented arm as well as providing a high contrast image. The grey background had an intensity of $\frac{32}{255}$ ($\approx \frac{1}{8}$ on a scale from black (0) to white (1)), which was selected instead of a black background because initial tests showed poor optimisation performance in these conditions.

Instead of using the original realistic skin texture on the arm, a solid non-reflective near-white colour was used to prevent model optimisation problems resulting from having to learn the reflective properties of the arm with respect to the environment. This near-white colour had an intensity of $\frac{232}{255}$ ($\approx \frac{7}{8}$ on a scale from black (0) to white (1)), which was chosen for the same reason optimisation reasons as the grey background colour. Finally, the ML-Agents toolkit was used to provide the interface between the Unity environment and the model in Python [30].

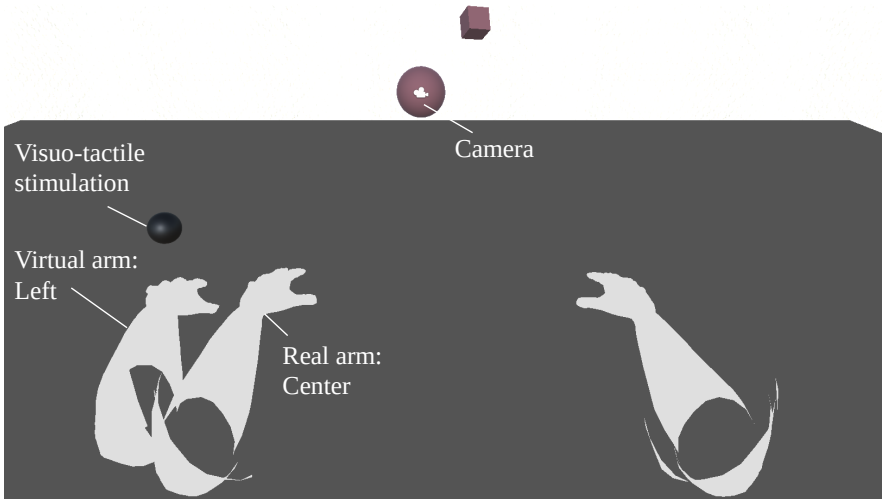


Fig. 3: Virtual environment and experimental setup modelled in the Unity engine.

3) Rubber-hand illusion: Equivalent to the setup used by [23], the real arm was placed in a forward resting position such that the hand was located 30 centimeters to the left of the body midline, denoted as the center position. The real arm was not visually observed by the agent, instead, the fake 'virtual' arm is observed either at the position of the real hand, 15 centimeters to the left of the real hand or 15 centimeters to right of the real hand, respectively known as the center, left and right positions. These conditions were created in the environment by manually adjusting the shoulder and elbow joint to making sure that the horizontal (X-axis) distance between the center and left/right condition was 15 centimeters, and the vertical (Y-axis) distance was zero centimeters. Because of the limitation of only being able to control two joint angles, this resulted in a slight deviation in the forward/backward (Z-axis) distance of about 1 centimeter, making the absolute distance between the left/right and center conditions a few millimeters higher than 15 centimeters.

Visuo-tactile stimulation was applied by generating a visual event at a regular interval of two seconds, followed by tactile event after a delay randomly sampled from a uniform distribution within a certain range. This range was depended on the stimulation condition, being either synchronous or asynchronous. In the synchronous condition, the time difference (in seconds) between the visual and tactile event was uniformly sampled from the range [0, 0.1), while for the asynchronous condition, it was sampled from the range [0, 1). The rate of exponential decay when no event took place r_{decay} was set to $\frac{1}{80}$. The variance of the Gaussian distribution determining the likelihood of the visual and tactile events shown in Equation (12) was set to 0.2 (with t_v and t_t in seconds).

The agent's arm was restricted such that no actual movement of the real arm was allowed, instead, the attempted movement was being observed. For both visual models, the agent was exposed to the three RHI conditions for

$N = 5$ trials. Every trial lasted for 1500 iterations, which in combination with $\Delta t = 0.02$ means an analogous real trial time of 30 seconds.

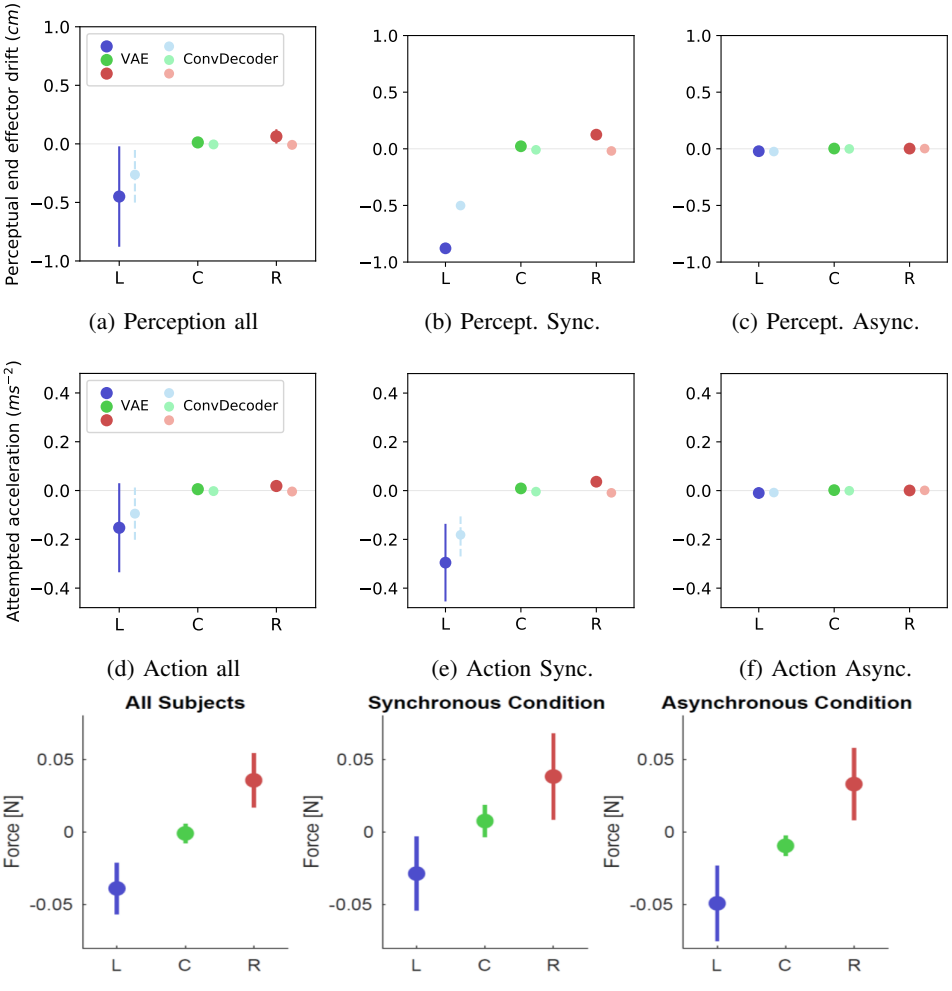
4) Reaching: To validate proper functioning of the agent's perception and action, multiple reaching tasks were performed, testing the agent's ability to move its arm to a desired position. To this extend, 80 random goal arm positions were generated, of which 40 were close to the initial arm position and 40 were further away from the initial arm position. For the 'close' positions, joint angle offsets for the shoulder and elbow were randomly sampled from $\pm\ Uniform(1, 8)$, while for the 'far' positions, these offsets were sampled from $\pm\ Uniform(8, 18)$.

In agent was informed of these goal positions by providing it the visual perception corresponding to the goal positions as a visual goal in the active inference formulation. Both visual models were subjected to the reaching task using the same goal positions for 5000 iterations per reaching task. Both the distance between the real arm and the goal position and the visual error were recorded and are shown in Fig. 7.

III. Results

A. Rubber-hand illusion

The shift in perceived arm location and the action attempted by the agent resulting from subjection to the RHI are depicted in Fig. 4. Both showed similar patterns for both the synchronous and asynchronous condition. We take



(a) Perception all (b) Percept. Sync. (c) Percept. Async.

(d) Action all (e) Action Sync. (f) Action Async.

(g) Human recorded action (forces)

Fig. 4: Model results. (a, b, c) Mean perceptual end-effector drift (in cm). (d,e,f) Mean attempted horizontal end-effector acceleration. (g) Mean forces exerted by human participants in a virtual rubber-hand experiment (from [23]). (h) Visual representation of the Jacobian learnt for the visual models.

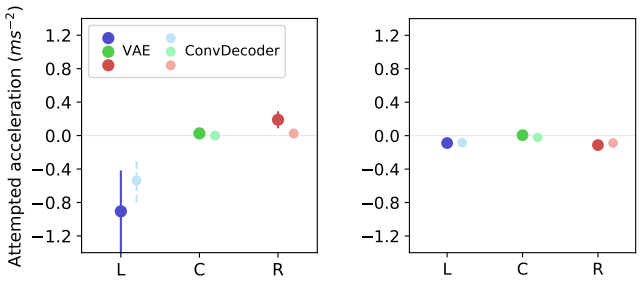
	VAE		ConvDecoder	
Condition	μ	σ	μ	σ
Left	-2.954×10^{-1}	1.592×10^{-1}	-1.815×10^{-1}	8.857×10^{-2}
Center	9.011×10^{-3}	2.939×10^{-3}	-4.063×10^{-3}	1.806×10^{-3}
Right	3.649×10^{-2}	2.034×10^{-2}	-9.099×10^{-3}	6.531×10^{-3}

(a) Synchronous stimulation

	VAE		ConvDecoder	
Condition	μ	σ	μ	σ
Left	-9.7594×10^{-3}	2.254×10^{-2}	-8.012×10^{-3}	7.392×10^{-3}
Center	1.957×10^{-3}	1.827×10^{-3}	-6.222×10^{-4}	1.108×10^{-3}
Right	5.990×10^{-4}	4.253×10^{-4}	1.015×10^{-3}	7.207×10^{-4}

(b) Asynchronous stimulation

Fig. 5: Statistics of the attempted horizontal end effector acceleration under the RHI (in ms^{-2}). The numbers shown are analogous to the results displayed in Fig. 4e and 4f.

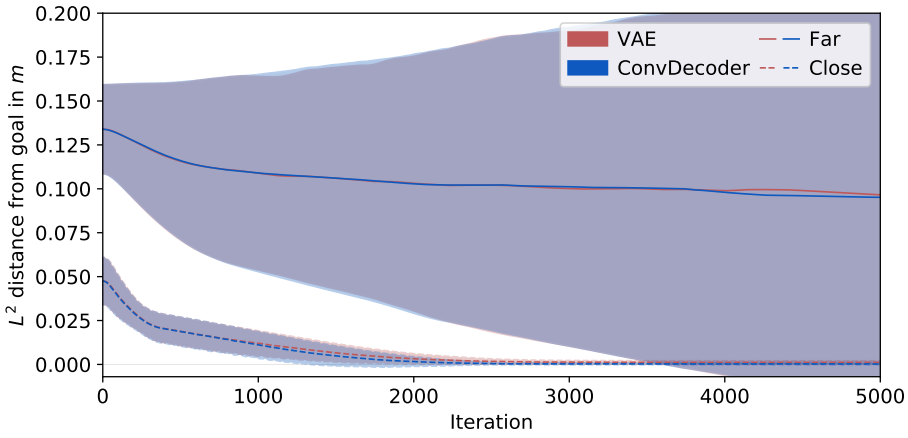


(a) Shoulder joint (b) Elbow joint

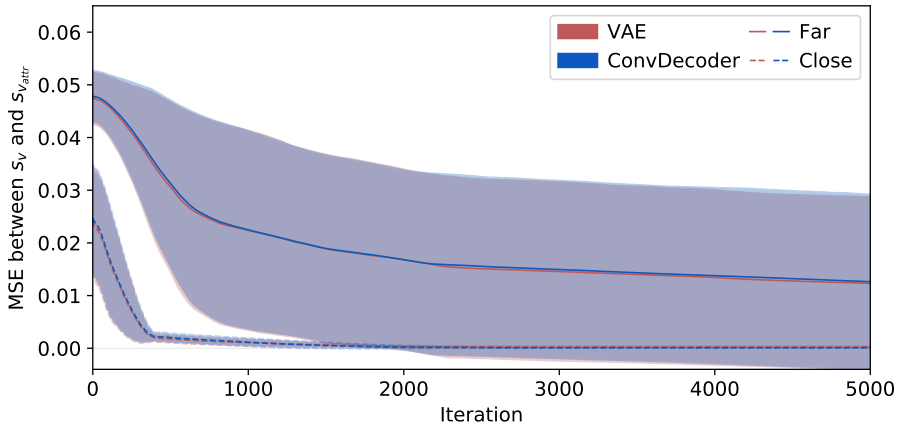
Fig. 6: Attempted joint angle acceleration under the RHI during synchronous stimulation.

a more in-depth look at the action patterns, since these are the results to be compared with the human data. For the left condition, we observed the expected negative (i.e. left on the horizontal axis) end effector action during synchronous stimulation (4e) for both visual models. For the center condition, both models seemed to produce the expected near-zero end effector action. For the right condition, we observed the expected positive end effector action for the agent using the variational autoencoder during synchronous stimulation, but a close to zero action for the convolutional decoder. It seems to be the case that the convolutional decoder in general produced weaker actions than the variational autoencoder under the same conditions. Lastly, we found that during asynchronous stimulation, both models produce near-zero action under any experimental condition. It is interesting to note that for both models, left-oriented action was multiple orders stronger than right oriented action (for the VAE under synchronous stimulation, the ratio was $\frac{\mu_L}{\mu_R} \approx 8.1$), even though these conditions had identical distances from the center condition.

To better understand the attempted end effector acceleration we examine the components from which it originates, being the attempted acceleration in the shoulder and elbow joints which is displayed in Fig. 6. For the shoulder joint, we observed similar action patterns for both models as observed in the end effector. However, the action of the shoulder joint for the right condition appeared to be stronger (relative to the action for the left condition) than in the end effector for both models. We identify the origin of this difference by looking at the action in the elbow joint. For the elbow joint we observed an attempted downward rotation for the left and right conditions. Because the shoulder joint was positioned at angle of about 30° relative to the vertical axis, downward rotation had a left horizontal component while upward rotation had a right horizontal component. This means that the elbow joint produced an attempted action with a left horizontal component for both the left and right conditions. Subsequently, the expected right horizontal component produced by the shoulder joint was hindered by left horizontal component produced by the shoulder joint, increasing the difference between the left and right condition.



(a)



(b)

Fig. 7: Reaching performance. (a) Euclidean distance from the middle of the agent's hand to the middle of the target hand position. (b) Mean squared error (MSE) between the visual perception s_v and the visual goal $s_{v_{goal}}$.

B. Reaching

For the reaching tasks we measured both the L^2 distance between the real arm to the goal and the mean squared error (MSE) between the visual observation and the visual goal, shown in Fig. 7. We observed similar performance for the variational autoencoder and the convolutional decoder models for both the 'close' and 'far' reaching tasks. For the 'close' reaching tasks, we observed quick and stable convergence to the goal position, where the per-task deviation was mostly caused by the random starting position. For the 'far' reaching tasks on the other hand, the average of all runs showed only a slight trend towards the goal position, with a larger variance than expected from the random initial start position only. We found that there were twelve goal positions for which both models were not able to decrease the distance between the end effector and the goal, indicating a complete failure to reach towards the target. These goal positions all had their shoulder joint angle within [-18, -15.3], with eight positions having their elbow range within [-17.7, -8.3] and the other four positions within [10.9, 18]. Analysing the data without these outliers (Fig. 8) shows a stable trend for the 'far' condition similar to that observed in the 'close' condition, indicating that the overall 'far' reaching performance was adequate, but that a certain range of goal targets pose a problem to the model.

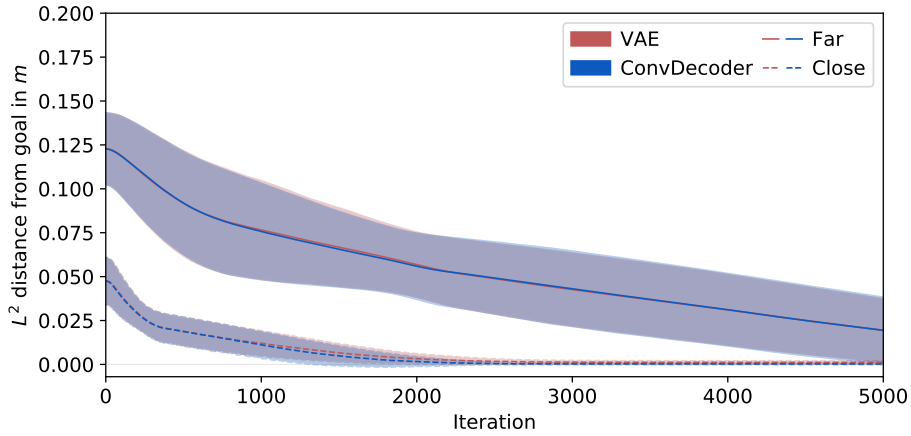
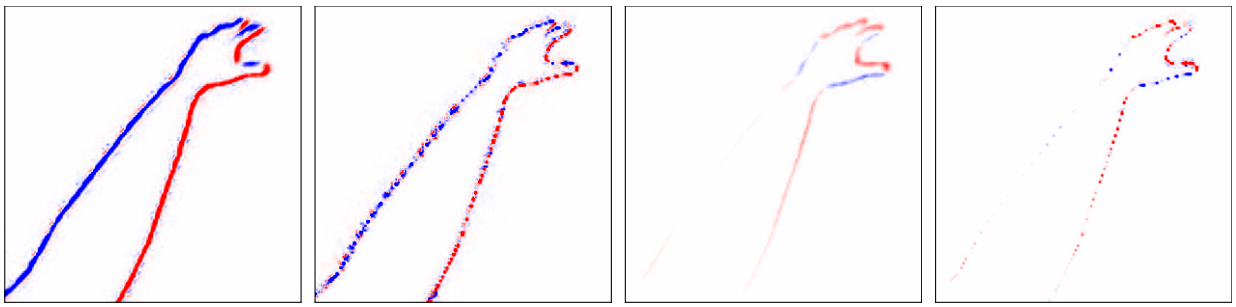


Fig. 8: Euclidean distance from the middle of the agent's hand to the middle of the target hand position without the twelve outliers discussed in section III-B.



(a) VAE: Shoulder (b) ConvDecoder: Shoulder (c) VAE: Elbow (d) ConvDecoder: Elbow

Fig. 9: Visual representation of the $\frac{\partial g(\mu)}{\partial \mu}$ Jacobian for both visual models. Positive (red) points indicate that increasing the corresponding joint angle increases the intensity of those pixels, while negative (blue) points indicate a decrease in pixel intensity for an increased joint angle.

IV. Discussion

A. Agent performance

We have seen how the free-energy agent produced action roughly within our expectations of its behaviour under the RHI. In addition, we have seen that the model is well capable of reaching towards a goal position presented in the form of a visual goal, except for a certain range of goal positions which the agent was unable to approach. The fact that all outliers consistently arose from goal positions in a very specific range of joint angles suggests that this range of positions is associated with local optima in the visual model. This could be partly caused by the arm moving out of the camera frame, but because this issue not only appears for highly negative goal elbow joint angles, but also for highly positive goal elbow joint angles, it is unlikely to be the main cause.

For the RHI trials, the results showed that even though the left and right conditions are located at exactly the same distance from the center position, action towards the left was stronger than action towards the right. Although non-symmetry of the actions could be explained by phenomena such as the body midline bias, these phenomena generally expect a stronger illusion (or action, see [23]) when the rubber arm is closer to the body, meaning that we would expect to see a stronger action for the right condition with respect to the left condition. We also observed that the shoulder joint produced (downward-)left oriented action for both conditions, and that the elbow joint action was weaker than the shoulder joint action.

To understand how these action patterns arise, we examine at the Jacobian $\frac{\partial g(\mu)}{\partial \mu}$ of the visual models. We have discussed how this term, shown in Fig. 9, constitutes the mapping between joint angle space and visual space, which allows us to interpret what happens during the backward pass through the network. The mapping is intuitive,

showing how the pixel intensity values of the visual prediction $g(\mu)$ are expected to change with respect to a change in the joint rotation. As to be expected, the shoulder Jacobian has positive values on the right and negative values on the left outline of the arm, as positive shoulder rotation will cause the arm to move to the right. For the elbow joint, we see that the top outline contains positive values and the bottom outline negative values, as positive elbow rotation causes the arm to move up and towards the camera.

From the visualisation of $\frac{\partial g(\mu)}{\partial \mu}$, we take away three important conclusions. First of all, the variational autoencoder shows more consistent and less noisy Jacobians, indicating that this model might have learned a more stable representation of the visual space and its relation to the joint angle space.

Secondly, we observe that the values of the elbow joint Jacobians are weaker than the shoulder joint Jacobians for both models. We hypothesise that this is the reason why we have observed weaker elbow joint action under the RHI, as well as it being able to explain why elbow joint convergence during the reaching tasks was slower than shoulder joint convergence.

Finally, conversion from visual space to joint angle space greatly emphasises local differences. For both models and joints the strongest values lie on the border of the arm. This means that when the visual prediction error is multiplied with the Jacobian to obtain the visual belief and action update, the error at the border of the arm is of the greatest influence. By combining this knowledge with the visual location of each of the RHI conditions (Fig. 10), we find a possible explanation for the asymmetric action patterns. We observe that the visual representation of the left condition overlaps with the visual representation of the center condition, while the visual representation of the right condition, this is not the case. We therefore hypothesise that the local emphasis exerted in the error minimisation combined with the visual perspective are the cause of the difference in action strength between the left and right conditions.

In addition, the local emphasis in the visual space also fits the reaching results, as it could explain why goal positions that are visually far away from the initial position are hard to reach than goal positions that are visually close to the initial position.

Fig. 10: Visual observation of the RHI conditions.

B. Comparison with human data

Our findings resemble the human data from [23] (shown in Fig. 4g) in that a non-negative action is produced in the direction of the rubber arm. Our results deviate from these findings in that we observe asymmetric action for the left and right condition despite their equal distance from the hand. In addition, under asynchronous stimulation our model produces close to no action, while the human data shows no statistically significant difference between the forces produced under synchronous and asynchronous stimulation [23]. This indicates that the prior probability value for γ (and/or general visual component strength) was set too low, where higher values might have attained similar results to that in humans.

Aside from comparing the patterns shown, in both human and agent data on shared characteristics, we could not perform direct statistical analysis between the two. Because the agent's action was simplified to velocity control and thereby ignored many physical influences (see section IV-C) in the system, we were measuring attempted acceleration instead of end effector force measured in humans [23]. Although these metrics are similar in meaning, the simplifications made to the environment prevented us from being able to perform direct comparison.

C. Model simplifications and challenges

In order to create a working system that produced valuable results within the time constraint, a number of simplifications had to be made.

First of all, the active inference formulation that drove the agent's action was build upon extensive simplifications such as the Laplace and mean-field approximations, in which we simplify the approximation of a probability distribution to the approximation of its mean. Although easy in computation and interpretation, this ignores the information captured by other properties of the distribution. Better results could possibly be achieved by using variational methods that do not make these approximations (see for example [31]).

Secondly, the visual system was simplified to use monocular grayscale vision instead of binocular RGB vision in order to make visual model training feasible. Although we speculate that colour vision would not have a great effect on model performance, binocular vision might have a significant effect as at would allow for three-dimensional interpretation of the image. This could possibly alleviate some of the optimisation problems discussed in section IV-A which are likely due to the error computation in two-dimensional visual space. However, this would come at the cost of challenging model training and a diminished real-time execution performance. Another approach would be to use monocular vision but to perform the error computation at multiple levels in the network, allowing for error computation on multiple abstract space representations. Such an approach corresponds to formulations of the free-energy principle that model perceptual inference as an hierarchical error optimisation in the brain [9].

The background and arms were solidly coloured to ensure proper optimisation and to rule out any influence from the environment on model performance. Although the lack of influence from external factors in the visual space is desired from a scientific and proof-of-concept perspective, real-world applications would need the visual model to account for noise and visual distractors in order to function properly[3].

Important simplifications made to the environment and free-energy model include the exclusion of mechanics such elasticity, angular and end effector drag and mass. As mentioned before, the model was simplified to act upon a velocity controller in the joint angles, which allowed for relatively simple environment setup and model execution. In addition, only considering the first order dynamics in the free-energy formulation allowed for its simple implementation, execution and parameter configuration.

A disadvantage associated with these simplifications is that we could not directly compare the agent's results with the data produced by humans from [23], although we argue that configuring a complex physics environment such that it can accurately represent the real world experimental setup (and corresponding empirical results) would be infeasible within the given time constraint.

By using a simple time-step update to compute μ and a, we are in fact using a simple proportional controller. Such simple controllers are known to oscillate for high gains and experience slow convergence for lower gains. This means that by including higher order dynamics in the model's formulation, a substantial improvement in reaching speed and convergence performance could possibly be achieved.

The usage of a simple time-step update also meant that when there is a persistent error, for instance when the real arm is restricted and cannot move to minimise the error, the action will increase indefinitely. This is not biologically plausible as at some point, the neuronal activation driving the action is saturated and cannot increase further. To account for this, the action value was clamped to simulate this saturation.

Finally, the environmental setup included no delay between the occurrence of events and their perception by the agent. This lack of sensory delay meant that we did not have to worry about the model and real world being out of sync possibly causing additional optimisation challenges, as shown in [16]. However, in both humans and robotic implementations, this delay is inherent to the system and thus unavoidable.

D. Future work

The results have shown how the agent produced action similar to humans under the RHI, as well as showing adequate reaching performance. However, limitations in the visual model caused sub-optimal performance under the RHI and in reaching tasks. It would be interesting for future work to focus on visual architectures that improve perception and action updating by instead of performing the error computation purely in the two-dimensional visual space, perform error computation and propagation at multiple levels as described in [32]. This could allow for a

[3]The problem with bringing this model to the real world is that generating a correct visual prediction for a real-world environment requires intricate knowledge about that environment. This problem can be solved by either only performing visual prediction and error computation for the interesting part of the visual space (say, the visual prediction of a limb), or by creating a universal model that incorporates the world knowledge and reasoning necessary to predict the visual result real-world dynamics. Note that if we were able to create such a model, the agent would not only be able to update beliefs about their limb positioning but also about events and properties in the environment.

more global and perspective invariable visual system as it is not limited by the visual space, instead being able to perform error computation at multiple abstract layers of the visual system. Furthermore, using less approximations in the active inference formulation (e.g. by using stochastics methods [31]) will possibly result in more accurate and complete action and perception.

Additionally, it is important to note that our implementation used static values for controlling variables in the active inference formulas such as the precision of the sensory modalities Σ_v^{-1} and Σ_p^{-1}, while in fact, these should be optimised in real-time depending on the perceived reliability of the modalities. The optimisation of these terms in real-time would enable the model to dynamically change the influence of sensory modalities, allowing the model to adjust for noisy sensory input increasing precision. In addition, future work could focus on extending the active inference formulation of the model by adding physical mechanics and replace the velocity control by force actuation. This could allow the model to be used in more sophisticated simulations and possibly even real-world applications. Finally, in this work, we only consider pure reactive action to prediction error. It would be interesting to consider models of action with policy, where high-level planning is involved and a policy is learned based on the task.

E. Conclusion

In this work, we have proposed a deep active inference model with as goal to replicate the action patterns found in [23]. The model used high-fidelity visual input in combination with proprioception to guide perception and action in a simulated environment. The model was able to produce similar perceptual and force patterns to those found in humans, reinforcing active inference as a viable method to model human action. Additionally, reaching tasks validating the proper function of the agent's action by successfully reaching targets both close to and far from the initial position, showing no difference in model performance. However, multiple outliers during the far reaching tasks showed that the visual model is likely susceptible to local optima which negatively impact the reaching performance. Possible future work includes the extension of the visual model and active-inference formulation to more precisely update the perception and action, as well as considering models that do not only depend on purely reactive action.

REFERENCES

[1] Konrad P. Körding and Daniel M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, Jan 2004.

[2] David C. Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712 – 719, 2004.

[3] Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, Jan 2002.

[4] Xu F. Word learning as bayesian inference. *Psychological Review*, 114(2):245, 2007.

[5] Alan A. Stocker and Eero P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585, Apr 2006.

[6] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, Oct 2000.

[7] Hermann von Helmholtz. *Handbuch der physiologischen Optik*. Leopold Voss, Leipzig, 1867.

[8] Karl Friston. A theory of cortical responses. *Philosophical Transactions: Biological Sciences*, 360(1456):815–836, 2005.

[9] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1):70 – 87, 2006. Theoretical and Computational Neuroscience: Understanding Brain Functions.

[10] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, Feb 2010.

[11] Karl J. Friston, Jean Daunizeau, James Kilner, and Stefan J. Kiebel. Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102(3):227–260, Mar 2010.

[12] K. Gu, G. Zhai, X. Yang, and W. Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, 2015.

[13] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, 21(1):41–52, 2012.

[14] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi. Just noticeable difference estimation for images with free-energy principle. *IEEE Transactions on Multimedia*, 15(7):1705–1710, 2013.

[15] Pablo Lanillos and Gordon Cheng. Adaptive robot body learning and estimation through predictive coding. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4083–4090. IEEE, 2018.

[16] Cansu Sancaktar, Marcel van Gerven, and Pablo Lanillos. End-to-end pixel-based deep active inference for body perception and action. *arXiv preprint arXiv:2001.05847*, 2020.

[17] Pablo Lanillos and Gordon Cheng. Active inference with function learning for robot body perception. *International Workshop on Continual Unsupervised Sensorimotor Learning, IEEE Developmental Learning and Epigenetic Robotics (ICDL-Epirob)*, 2018.

[18] Matthew Botvinick and Jonathan Cohen. Rubber hands 'feel' touch that eyes see. *Nature*, 391(6669):756–756, Feb 1998.

[19] Andreas Kalckert and H. H. Ehrsson. The onset time of the ownership sensation in the moving rubber hand illusion. *Frontiers in Psychology*, 8:344, 2017.

[20] Xaver Fuchs, Martin Riemer, Martin Diers, Herta Flor, and Jörg Trojan. Perceptual drifts of real and artificial limbs in the rubber hand illusion. *Scientific Reports*, 6(1):24362, Apr 2016.

[21] Majed Samad, Albert Jin Chung, and Ladan Shams. Perception of body ownership is driven by bayesian sensory inference. *PloS one*, 10(2):e0117178–e0117178, Feb 2015.

[22] Nina-Alisa Hinz, Pablo Lanillos, Hermann Mueller, and Gordon Cheng. Drifting perceptual patterns suggest prediction errors fusion rather than hypothesis selection: replicating the rubber-hand illusion on a robot. *arXiv preprint arXiv:1806.06809*, 2018.

[23] Pablo Lanillos, Sae Franklin, and David W. Franklin. The predictive brain in action: Involuntary actions reduce body prediction errors. *bioRxiv*, 2020.

[24] Friston Karl. A free energy principle for biological systems. *Entropy (Basel, Switzerland)*, 14(11):2100–2121, Nov 2012. 23204829[pmid].

[25] Karl Friston, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. Variational free energy and the laplace approximation. *Neuroimage*, 34(1):220–234, 2007.

[26] Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017.

[27] Karl Friston. What is optimal about motor control? *Neuron*, 72(3):488 – 498, 2011.

[28] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks, 2014.

[29] Carl Doersch. Tutorial on variational autoencoders, 2016.

[30] Arthur Juliani, Vincent-Pierre Berges, Esh Vckay, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A general platform for intelligent agents, 2018.

[31] John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search, 2012.

[32] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293 – 301, 2009.