Modeling the Anticipatory Remapping of Spatial Body Representations: A Free Energy Approach

Patrick Weigert (patrick.weigert@student.uva.nl)
Institute for Logic, Language and Computation, Science Park 107, Amsterdam, 1098 XG Netherlands

Johannes Lohmann (johannes.lohmann@uni-tuebingen.de) &
Martin V. Butz (martin.butz@uni-tuebingen.de)
Neuro-cognitive modeling group, Department of Psychology and Department of Computer Science,
University of Tübingen, Sand 14, Tübingen, 72076 Germany

Abstract

According to theories of event-predictive cognition, neural processing focuses on the next relevant interaction targets. Evidence for this notion comes from the anticipatory crossmodal congruency effect (aCCE), which implies that spatial body representations are mapped onto future goal locations in advance of a goal-directed action. Here we present a free energy based normative process model that accounts for the aCCE quantitatively by applying crossmodal mappings between vision and touch as well as active inference. A comparison with a diffusion model shows that our model accounts for the response time distributions and the aCCE with a sparser set of parameters. However, the temporal dynamics of the model require further fine tuning to account for all aspects of the aCCE. The model shows how the free energy framework can be used to account for behavioral data in general and how to implement theories of event-predictive cognition in a normative cognitive process model.

Keywords: active inference; free energy; event-predictive cognition; crossmodal congruency

Introduction

Theories of event-predictive cognition suggest that the central nervous system develops internal generative event models, predicting its environment in order to facilitate the invocation and interpretation of goal-directed behavior (Butz, 2016; Kuperberg, 2020; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). Event models encode predictions about the event dynamics as well as start and end conditions, which characterize event boundaries. In neuroscience, the concept of generative models was introduced by Rao and Ballard (1999) in their predictive coding theory, which implements perception as an inference process, where top-down predictions disambiguate bottom-up sensory signals. Discrepancies between top-down predictions and sensory signals yield a prediction error that can be used to adapt the model. Meanwhile, active inference can invoke actions as the realization of top-down proprioceptive predictions via muscular activation (Adams, Shipp, & Friston, 2013). Together, predictive coding and active inference provide a general formalism to describe neural computations. Combined with event-predictive models, active inference does not only activate immediate sensory consequences, but also potentially anticipated event boundaries and subsequent events.

In this paper we introduce a variational active inference framework, which models response times from a behavioral experiment by means of a normative process model (Lewandowsky & Farrell, 2011). The experiment was originally designed to probe active inference on a behavioral level (Belardinelli, Lohmann, Farnè, & Butz, 2018; Lohmann, Belardinelli, & Butz, 2019). In particular, significant anticipatory cross-modal interactions between light and tactile stimulations show that our minds anticipate future bodily states relative to the surrounding environment. Our normative model implementation offers an algorithmic explanation of these interactions based on computational theories of active inference and event-predictive cognition.

Active Inference on a Behavioral Level

Eye-tracking studies allow to probe the information sampling processes unfolding during goal-directed actions. The observed gaze patterns imply an anticipatory mode of control, where the eyes are ahead of the hand and thus tend to fixate next relevant hand targets (Hayhoe, Shrivastava, Mruczek, & Pelz, 2003). More recently, Belardinelli, Stepper, and Butz (2016) have shown that the fixations while preparing to and grasping an object predict the type of grasp that is going to be applied. The eyes tended to look towards future index finger positions, presumably because the index finger was the first one to pass and then touch the object. This selective sampling of critical information regarding a certain sensorimotor prediction fits well with the assumptions of active inference. However, an exact algorithmic explanations of the involved inference processes and underlying representational formats is missing. Our model closes this explanatory gap.

The eyetracking results imply that the future hand and finger locations on the target object are predicted in advance of the actual action. Note, though, that this does not necessarily mean that the whole prediction process takes place in an eye-centered frame of reference or that the finger locations are predicted precisely. We hypothesize that coarse predictions are made relative to the effector and the target object. These predictions may then be mapped into an eye-centered frame of reference. Such crossmodal mappings are a core feature of certain spatial body representations, namely of peripersonal space (PPS; Fogassi et al., 1996).

In line with the definition of Bufacchi and Iannetti (2018), we consider PPS as a set of response fields, whose activation depends on the behavioral relevance of a certain stimulus. Hence, PPS is not primarily constrained by distance but by the relevance to interact with or to avoid a certain object. In order to realize this function, PPS provides spatial mappings between modalities, for instance between haptic perception and motor codes, allowing for a fast response to tactile sensations without looking. Direct mappings between sensory modalities, such as vision and touch, exist as well, yielding crossmodal congruency effects (CCE; Spence, Pavani, Maravita, & Holmes, 2004), where visual stimuli facilitate the detection of touch when presented close to the stimulated body surface (congruent condition), while they partially interfere when presented close to other relevant body parts (incongruent condition).

If PPS is indeed used to generate predictions and thus provide control signals regarding the future hand position, one would expect to observe, for example, a CCE at the future hand position. Such findings would not only support the idea that the predictive process observed in eyetracking studies is rooted in PPS. They would also support one of the central claims of event-predictive cognition in general, as this would indicate a focus of neural processing on the next relevant event boundary. Indeed, PPS has been found to be highly adaptive and to change with interaction possibilities and interaction goals. While modulations of PPS due to tool use are well established (cf. Bufacchi & Iannetti, 2018), recent findings show that PPS is remapped to the future hand position during goal-directed object interactions.

Anticipatory Crossmodal Congruency Effect

One of the first experiments indicating that PPS is indeed mapped onto the target of a goal-directed movement was reported by Brozzoli, Cardinali, Pavani, and Farnè (2010). Participants had to perform either a grasping or a pointing movement towards a cylinder. Before, at movement onset, or during the movement, participants received an electric stimulation at the thumb or index finger. At the same time, a LED was flashed at the target object, either at the approximate landing position of the thumb or index finger. Brozzoli et al. (2010) observed a CCE between the visual and the tactile stimulus. The advantage for congruent stimulation became stronger for later stimulations, i.e. the CCE was larger for stimulations during the movement than for stimulations before the movement. The authors argued that these results indicate an action-related remapping of PPS towards the target object. Seeing that the observed CCE occurred with respect to the future finger positions, we call it the anticipatory crossmodal congruency effect (aCCE). Several follow-up studies probed the properties of the aCCE. These results imply that the aCCE (i) is tied to movement planning and occurs on a trial-wise basis as soon as the target becomes visible (Belardinelli et al., 2018; Lohmann et al., 2019), (ii) becomes stronger for later stimulations (Belardinelli et al., 2018; Lohmann et al., 2019), (iii) is modulated by planning certainty, that is, it is reduced if the sensorimotor mapping becomes less predictable (Lohmann et al., 2019), and (iv) occurs mainly for the currently relevant movement target in a sequential interaction (Lohmann & Butz, 2020).

The aCCE supports theories of event predictive cognition in general, but so far there is no formal normative model to account for it. The quantitative model introduced here provides a free-energy based account for the anticipatory remapping of PPS, which is considered as an active inference process. Furthermore, the model also models the relation between the predictive state of the agent and the response to the tactile stimulation. Due to the assumed crossmodal mappings, the anticipatory remapping emerges due to the activation of future goal states by active inference. The next section describes the central model formalism. After introducing the model, we evaluate our model in comparison to a diffusion model, which we use as the baseline.

Model Description

In the following, we follow the Free Energy method as described by Friston (2010) and Buckley, Kim, McGregor, and Seth (2017). An overview of the model is shown in Fig. 1. The general idea is to model a reaching action in the light of its final goal state, that is, the grasp of the target object. In terms of event-predictive cognition, the final goal state can be encoded by an event boundary, which indicates the end of the particular event that generates the goal state (Zacks et al., 2007). In our simulations, the active inference process essentially invokes a goal-directed motion by the pre-activation of the desired goal posture (i.e. a grasp).

The environment of the agent is modeled as a multidimensional state variable ϑ. It consists of a hand and a bottle, both of which are located and rotated in space along one depth axis. Both, hand and bottle have a certain rotation along that axis. The index finger and the thumb, as well as the left and right grip points on the bottle are described using cylindrical coordinates with a fixed radius. Furthermore, the angles of index finger and thumb as well as left and right grip points are assumed to have a fixed difference of 180°. The considered environmental state can thus be characterized by

$$\vartheta = [x_h \quad x_b \quad \alpha_h \quad \alpha_b \quad s_t \quad s_v]^T ,$$

where x_h and x_b specify the positions of hand and bottle along the depth axis, and α_h, α_b their respective rotations. Quantities s_t and s_v encode the state of the tactile and the visual stimulation as binary variables (either index/thumb tactile stimulation or right/left visual distractor). Fig. 1 shows how the environmental reality is encoded into the state variables in ϑ.

The agent perceives the environment via sensory data φ. The process linking ϑ and φ is referred to as the *generative process* g, which is a (possibly non-linear) mapping of the state of the environment to the mean of a Gaussian distribution of sensory data,

$$\varphi \sim \mathcal{N}(g(\vartheta), \Pi_\varphi^{-1}),$$

where Π_φ is the precision (i.e. the inverse covariance) of the generated sensory data. The agent then tries to estimate the current state of the world $\widehat{\vartheta}$ given φ and its own internal model of the world, described by $\widehat{g}$ and $\widehat{\Pi}$. In our model, the agent

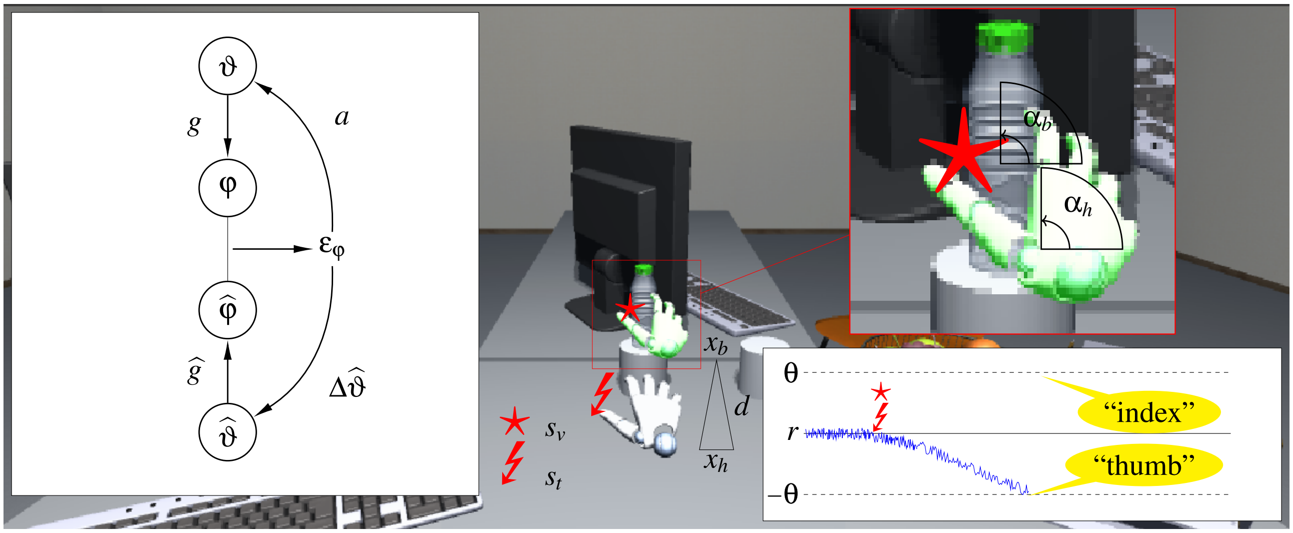

Figure 1: While reaching for a bottle, the distance $d = x_b - x_h$ reaches zero, while the angular orientation of the hand α_h is adjusted to the bottle's orientation α_b. At some point while the reaching behavior is initiated and executed, tactile stimulation s_t on one of the fingers and a concurrent visual distractor s_v are applied. In reaction to s_t the system models a verbal response r, which is influenced by the anticipated proximity of the stimulus s_v, yielding a compatible or incompatible reaction influence dependent on the bottle's orientation α_b and the side of the distractor s_v. The box on the left top shows how differences between expected and perceived sensory information trigger actions a (reaching motion) as well as adjustments of the internal environmental state estimations $\hat{\vartheta}$. The box on the bottom right depicts a schema of the diffusion model.

is allowed to perceive the world directly (with added noise), so $g(\vartheta) = \vartheta$.

The dynamics of the environment ϑ are described using a mapping f,

$$\vartheta' \sim \mathcal{N}(f(\vartheta, a), \Pi_{\vartheta'}^{-1}),$$

with an action variable a, which the agent invokes as motor activation. In our model, the agent can set the velocity of the hand using a_h and the velocity of the hand rotation a_α, essentially aiming at a particular target state a_h of the hand position x_h and a particular target hand orientation a_α for the hand's actual orientation α_h:

$$f\left(\begin{bmatrix} x_h \\ x_b \\ \alpha_h \\ \alpha_b \\ s_t \\ s_v \end{bmatrix}\right) = \begin{bmatrix} a_h \\ 0 \\ a_\alpha \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The generative model of the agent encodes the relationship between environmental states in a different fashion, namely via a hyperprior

$$\hat{\vartheta} \sim \mathcal{N}(h(\vartheta), \hat{\Pi}_{\vartheta}^{-1}),$$

where $\hat{\Pi}_{\vartheta}$ reflects the confidence in the assumed interdependence. Thereby we essentially model the hypothesized event-predictive active inference nature of the agent, who has the goal to align its hand position and orientation with the bottle. We thus encode the hyperprior by:

$$h\left(\begin{bmatrix} x_h \\ x_b \\ \alpha_h \\ \alpha_b \\ s_t \\ s_v \end{bmatrix}\right) = \begin{bmatrix} x_b \\ x_b \\ \alpha_b \\ \alpha_b \\ s_t \\ s_v \end{bmatrix}.$$

The agent then tries both to infer the true state of ϑ given its model of the world and to change the world to match its expectations via action. This is achieved by minimizing the Free Energy within the model. Because the above probability densities are Gaussian, the resulting Laplace approximation implies that minimizing the Free Energy involves minimizing the quantity (Friston, Trujillo-Barreto, & Daunizeau, 2008):

$$\frac{1}{2}(\varepsilon^T \hat{\Pi} \varepsilon - \ln|\hat{\Pi}|),$$

with a block diagonal precision matrix

$$\hat{\Pi} = \begin{bmatrix} \hat{\Pi}_{\varphi} & \\ & \hat{\Pi}_{\vartheta} \end{bmatrix},$$

and *prediction errors*

$$\varepsilon = \begin{bmatrix} \varepsilon_{\varphi} \\ \varepsilon_{\vartheta} \end{bmatrix} = \begin{bmatrix} \varphi - \hat{g}(\hat{\vartheta}) \\ \hat{\vartheta} - h(\hat{\vartheta}) \end{bmatrix}.$$

On this quantity, gradient descent is performed:

$$\Delta\hat{\vartheta} = \vartheta' - k_p \frac{\partial \varepsilon^T}{\partial \hat{\vartheta}} \hat{\Pi} \varepsilon,$$

with a learning rate k_p. In addition to adapting the estimated environmental state $\widehat{\vartheta}$ to minimize its prediction errors, the agent is assumed to have learned how its actions can change its sensory data. This knowledge is encoded in a quantity $\frac{\partial \varphi}{\partial a \partial t}$. The optimal action at any point in time is then given by

$$a = -k_a \int \frac{\partial \varepsilon_\varphi^T}{\partial a \partial t} \widehat{\Pi} \varepsilon_\varphi dt,$$

which essentially invokes hand motions towards the target object with the target orientation.

In the end, the agent is supposed to deliver a response indicating which finger was stimulated. Therefore, it can set a response variable r taking on possible values of s_t, and it should do so while integrating the estimated tactile and visual stimuli. This is assumed to happen in a Bayes-optimal fashion:

$$r = \frac{\widehat{s_t}\widehat{\Pi}^t_\varphi + \mu(\widehat{s_v})\widehat{\Pi}^v_\varphi}{\widehat{\Pi}^t_\varphi + \widehat{\Pi}^v_\varphi},$$

where $\widehat{\Pi}^t_\varphi$ is the tactile stimulus component of $\widehat{\Pi}_\varphi$, $\widehat{\Pi}^v_\varphi$ is the visual distractor component, and μ is a mapping representing the integration of the stimulus.

This integration process is driven by the crossmodal mappings inherent to peripersonal space. The strength of these mappings is affected by distance (cf. Spence et al., 2004), however, the exact relation between distance and integration strength remains open. For simplicity we assume a linear relationship, with no integration at the initial hand position and full integration at the target location,

$$\begin{aligned} \mu(s_v) &= k_i \operatorname{sgn}(s_v s_t)(1-d) \qquad (1) \\ d &= x_h - x_t, \end{aligned}$$

with an *integration coefficient* k_i reflecting the strength of the integration.

The response time is then determined as the duration between the appearance of the bottle stimulus and the point where r reaches a threshold θ, with an additional offset ρ as the *non-response time*. The evolution of r over time can be described as a random walk with normally distributed increments. Hence, it is similar to a Wiener process that is used to model response times in diffusion models.

To sum up, the model can describe changes in the hand position and orientation over time, as well as the response time for the tactile discrimination task.[1]

Results

To evaluate our model, we fitted the response time data from the second experiment reported in Lohmann et al. (2019). We focus here on the consistent condition with a predictable sensorimotor mapping. We fitted the data with a diffusion model and our free energy model and compared the goodness of fit of both approaches. The data set consists of 2926 response times from 24 experimental conditions. In the experiment we varied the bottle orientation (upright or upside down), the side of the visual distractor (on the left side of the bottle, or on the right side), the stimulated finger (index or thumb), and the time point at which the visual and tactile stimuli were applied (either after 250 ms after showing the bottle, at movement onset, or after 50% of the distance between starting position and bottle was covered). The onset time of the stimulation relative to the stimulus onset or movement onset is referred to as *stimulus onset asynchrony* (SOA). All of the 21 participants completed six trials, yielding up to 126 response times in total, per condition. Only trials with responses were considered. These data were used to fit the models. Thus, we trained both models on response time distributions across participants.

Diffusion model

To assess the quality of the free energy model, we fitted the data also with a diffusion model (Ratcliff, 1978). Diffusion processes can be used to model response times in binary decision tasks, like responding with thumb or index finger given a tactile stimulation in a crossmodal congruency task. A preference for either one of the two options is considered as a one-dimensional quantity that changes over time, until it reaches either one of two threshold values, referred to as decision boundaries. The change of evidence over time is described in terms of a Wiener diffusion process, where a systematic component yields a constant change of evidence over time, while an additional random component adds Gaussian noise.

Diffusion models allow to estimate response time distributions for binary decision tasks by sampling the diffusion process given a set of parameters. Different extensions of the simple diffusion model have been discussed (Wagenmakers, 2009). Here we use a basic version with four free parameters per condition: the interval between the decision boundaries (a); the strength and direction of the systematic component of the diffusion process, that is, the drift rate ν; the duration of a non-decision process contributing to the response time (t_0), which serves as a constant response time offset; and the starting value of the diffusion process (z), which biases the random component. For the optimization of the respective parameters we used the partial derivative method proposed by (Voss & Voss, 2007), and their *fast-dm* algorithm. We applied the model separately to each experimental condition, yielding four free parameters for each of the 24 conditions.[2].

After estimating the parameters, we randomly sampled 1000 values from the modeled distribution per condition and determined the KL divergences between the predicted and observed distributions. To estimate the KL divergence of two empirical density distributions, we used the algorithm from Pérez-Cruz (2008). The KL divergences ranged between 1.7 and 2.3 with an average of 2.0. Two examples

[1] A python implementation of the model can be found in the online supplementaries: `https://osf.io/p7zwn/`.

[2] The raw data, the estimated parameters, as well as the settings file for *fast-dm* can be found in the online supplementary, along with the R scripts used to generate the distributions (`https://osf.io/p7zwn/`).

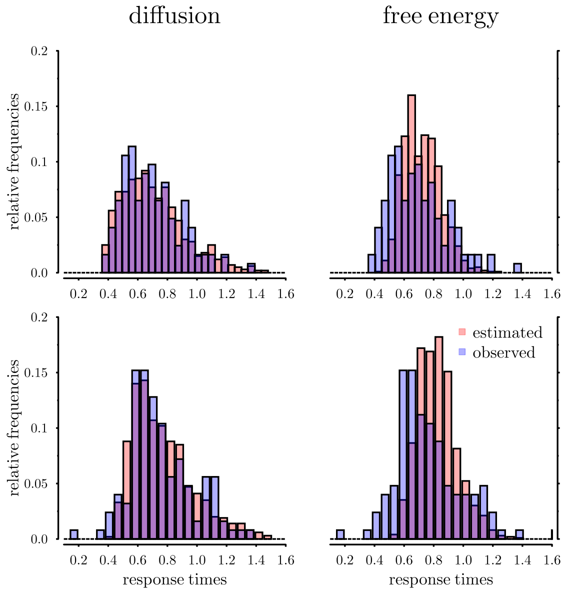

Figure 2: A comparison of fits provided by the diffusion model (left) and our free energy model (right). Red histograms show examples of the estimated distributions. The observed data is shown in blue. Data in the upper panels (congurent condition) was obtained with upright bottles, visual stimulation on the right side and tactile stimulation at the index finger during movement onset. Data in the lower panels (incongruent condition) was obtained with rotated bottles and otherwise the same stimulation setup. The free energy model was fitted to all conditions at once, yielding a less precise fit but still modeling the difference between the condition.

for estimated response time distributions are shown on the left side of Fig. 2. We estimated linear mixed models for the various factors and their combinations and performed a hierarchical model comparison. Only the SOA had a significant influence on the KL divergences. The respective Spearman correlation indicated a significant positive association between the ordered SOA levels (250 ms, movement onset, and during movement execution) and the KL divergences $(\mathrm{rs}(22) = .61, \mathrm{p} < .001)$. The KL divergences obtained with the diffusion model thus provides a performance reference for the free energy model. The reduced fit for later SOAs imply a systematic deviation from a typical diffusion process.

Free Energy model

The fit of the Free Energy model to the empirical data was carried out as follows. First, for simplicity, all precision matrices were assumed to be diagonal. The estimated precisions $\widehat{\Pi}^i_\varphi$ were set equal to the true precisions Π^i_φ, following the theoretical assumption that precisions are learnt by agents in a long-term learning process, i.e. longer than the duration of the experiment. Due to the structure of the h function in the model, out of the values in $\widehat{\Pi}_\vartheta$, only $\widehat{\Pi}^{x_h}_\vartheta$ and $\widehat{\Pi}^{\alpha_h}_\vartheta$ have an influence on the update equations. Therefore, the remaining components of $\widehat{\Pi}_\vartheta$ can be set to 0. The values for the learning rates $k_p = .1$, $k_a = .5$ were determined such that the simulation would be numerically stable. Similarly, the precision parameters involved in the grasping task were chosen so that the hand of the agent performs a natural motion. The remaining five parameters (visual and tactile stimulus perception, k_i, θ and ρ) were fitted for all the 24 conditions at once using a *Covariance matrix adaptation evolution strategy* (CMA-ES Hansen, 2006). We simulated 100 trials for each condition, determined the KL divergence between simulated and empirical data, and added up the squares of all divergences.

The KL divergences were more broadly distributed than for the diffusion model and ranged between 1.07 and 9.61 with an average of 5.09. Two examples for estimated response time distributions are shown on the right side of Fig. 2. Regarding the aCCE, the free energy model could reproduce its general magnitude, but the predicted SOA variation of the congruency effect was smaller than in the observed data (Fig. 3). Most likely this is due to the assumed linearity in (1), which should be investigated further.

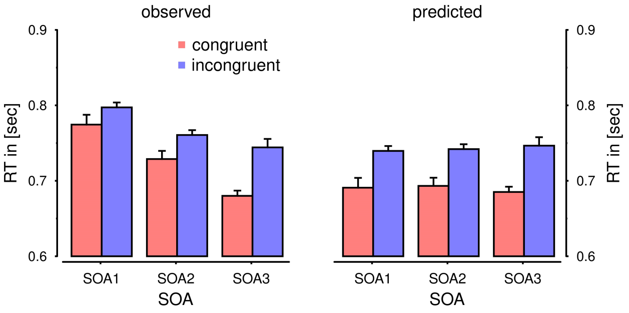

Figure 3: Observed (data taken from Lohmann et al., 2019) and predicted aCCE in the respective congruent (red bars) and incongruent (blue bars) conditions. Error bars show the standard error of the mean. The free energy model reproduces the congruency effect and its increase in strength qualitatively.

As we did for the diffusion model, we analyzed the KL divergence variations between the experimental factors by estimating linear mixed models for the various factors and their combinations and performed a hierarchical model comparison. Again, only the SOA had a significant influence on the KL divergences. The respective Spearman correlation indicated a significant positive association between the ordered SOA levels (250 ms, movement onset, and during movement execution) and the KL divergences $(\mathrm{rs}(22) = .47, \mathrm{p} = .019)$. Especially for stimulation at movement onset, the KL divergences were larger.

Like the diffusion model, our free energy approach had difficulties to account for responses at later SOAs. Compared to the diffusion model, the KL divergences show more variation in general and are overall higher, however, the free energy approach uses much less parameters.

Comparison

While the diffusion model is agnostic regarding the actual experimental setup, the free energy approach tries to capture core processes of the setup, like the assumed remapping of peripersonal space. This yields a lower number of free parameters compared to the diffusion process, however, fitting our free energy model to all conditions at once, led to systematic deviations, most notably for conditions with stimulation at movement onset. Given the sum of KL divergences in all conditions, one can obtain an upper bound on the Bayesian information criterion (BIC) score for both models using Jensen's inequality (Needham, 1993) as

$$\mathrm{BIC} \leq p \ln n \leq 2 \sum \mathrm{KL} + p \ln n,$$

with p as the number of free parameters (96 for the diffusion model, 12 for the free energy model) and n as the number of observations (in this case the 2926 response times in the 24 conditions). Due to smaller number of parameters the according BIC value for the free energy model is 340, compared to 961 for the diffusion model.

Discussion

Theories of event-predictive cognition and active inference essentially hypothesize that action effects and final event outcomes guide behavior (Butz, 2016; Friston, 2010; Kuperberg, 2020; Zacks et al., 2007). We introduced an event-predictive, free energy-based normative model, which is able to model behavioral dynamics. In particular, we focused on simulating behavioral response time data generated by means of an anticipatory crossmodal congruency effect (aCCE) paradigm. To elicit an aCCE, participants are asked to interact with an object and to concurrently report tactile finger stimulations as fast as possible (Brozzoli et al., 2010; Lohmann et al., 2019). While preparing and executing the action, light stimuli at the target object systematically interact with tactile stimulation on the hand before the hand actually reaches the object. In our model we assume that the agent predicts its future hand location before starting the action. This prediction is assumed to be realized by remapping peripersonal space to the future hand location. By means of free energy minimization the agent then tries to align its current hand position and orientation with the predicted one through action. The generation of the verbal response is modeled by a diffusion process that accumulates the evidence for a stimulation of either thumb or index finger. We assume an automatic, mandatory integration of visual and tactile stimuli within peripersonal space. Hence, the remapping of peripersonal space yields an integration of vision and touch at the goal location. This effect is assumed to become stronger the closer the hand is to the goal location. The results show that our model implementation elicits the aCCE across the different conditions, including a slight increase in strength for later SOAs—where the hand is closer to the bottle—in qualitative accordance with available data.

We contrasted the performance of our model with a diffusion process model, which modeled the different conditions separately, and thus the aCCE patterns from the data directly. This yielded a lower bound for the KL divergence, which can be obtained when fitting the data with a diffusion process. Meanwhile, the diffusion model has many more free parameters compared to our free energy model. A BIC-based model comparison confirmed that our event-predictive free energy model yields a better fit. Our model indeed produces a consistent congruency effect over the different SOA conditions. However, compared with the observed data, the increase of the congruency effect for later SOAs is not captured well. While our model predicts a larger difference between response times in congruent and incongruent conditions at later SOAs, the magnitude of this effect is much smaller than in the observed data. The reason for this mismatch is probably twofold. First, the variance in the observed data is higher for later SOAs. Since the KL divergence is asymmetric and punishes distributions that are too wide more than distributions that are too narrow, the KL-based optimization tends to produce distributions that underestimate the variance. Second, the discrepancy indicates that the assumption of a linear relationship between distance and the strength of the crossmodal mapping is not warranted. However, this relationship can be probed on a behavioral level and according data will help to improve the model.

While the results seem promising, showing that our free energy model can account for the data with a much sparser set of parameters than a saturated diffusion model, our study also indicates limitations in the current model implementation. Most notably, the model has difficulties to capture the magnitude of the observed variation of the aCCE over SOAs. This implies that the temporal dynamics of the modeled congruency effect require further fine tuning. Furthermore, the integration mapping μ, which plays a crucial role in modeling the dynamics of the effect, was designed by hand. For a pure free-energy account, further work would be required to derive this mapping from the underlying FEP framework instead. Finally, both models had difficulties to account for response times obtained at later SOAs. This might be due to the increased variance in response times for later SOAs. A closer examination or a computational explanation about how this data pattern comes about is pending. The assumptions of the free-energy model are yet to be empirically investigated. In particular, it remains unclear if the modeled direct access to the real world obscures additional behavioral effects when facing real visual data. The model also provides predictions with respect to the dependence of the aCCE on the precision parameters, which may be studied in an experimental setup where the variance of the sensory information can be altered. A virtual reality environment seems especially suitable to this end, as visual noise can be flexibly applied to the grasping target before or during the action.

Overall, the proposed model allows to simulate response times from unrolling an active inference process in a variational predictive model. This is a promising first step in applying the free energy framework to model actual behavioral data. In a next step we plan to model further data regarding the aCCE. For instance, we have shown that the aCCE is selectively reduced at movement onset if an unexpected mismatch between seen and felt hand movement is introduced (second experiment in Lohmann et al., 2019). Our model should be able to reproduce this finding by variations in the prediction errors. Improving our models ability to account for the temporal dynamics of the aCCE and probing its predictions on a behavioral level, will provide a useful tool to investigate the predictions of active inference in general, and event-predictive cognition in particular.

Acknowledgements

Funding was received from the German Research Foundation (DFG) (research unit FOR 2718 on Modal and Amodal Cognition, BU 1335/12-1, AOBJ: 661816). Martin Butz is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, *218*(3), 611–643. doi: 10.1007/s00429-012-0475-5

Belardinelli, A., Lohmann, J. Y., Farnè, A., & Butz, M. V. (2018). Mental space maps into the future. *Cognition*, *176*, 65–73. doi: 10.1016/j.cognition.2018.03.007

Belardinelli, A., Stepper, M. Y., & Butz, M. V. (2016). It's in the eyes: Planning precise manual actions before execution. *Journal of Vision*, *16*. doi: 10.1167/16.1.18

Brozzoli, C., Cardinali, L., Pavani, F., & Farnè, A. (2010). Action-specific remapping of peripersonal space. *Neuropsychologia*, *48*(3), 796–802. doi: 10.1016/j.neuropsychologia.2009.10.009

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79. doi: 10.1016/j.jmp.2017.09.004

Bufacchi, R. J., & Iannetti, G. D. (2018). An action field theory of peripersonal space. *Trends in Cognitive Sciences*, *22*(12), 1076–1090. doi: 10.1016/j.tics.2018.09.004

Butz, M. V. (2016). Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology*, *7*(925). doi: 10.3389/fpsyg.2016.00925

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., & Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area f4). *Journal of Neurophysiology*, *76*, 141–157. doi: 10.1152/jn.1996.76.1.141

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. doi: 10.1038/nrn2787

Friston, K., Trujillo-Barreto, N., & Daunizeau, J. (2008). Dem: a variational treatment of dynamic systems. *Neuroimage*, *41*(3), 849–885. doi: 10.1016/j.neuroimage.2008.02.054

Hansen, N. (2006). The cma evolution strategy: A comparing review. In J. Lozano, I. Larrañaga P. Inza, & E. Bengoetxea (Eds.), *Towards a new evolutionary computation. studies in fuzziness and soft computing* (pp. 75–102). Berlin, Heidelberg: Springer. doi: 10.1007/3-540-32494-1_4

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), doi: 10.1167/3.1.6.

Kuperberg, G. R. (2020). Tea with milk? a hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science*. doi: 10.1111/tops.12518

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks: Sage Publications.

Lohmann, J., Belardinelli, A., & Butz, M. V. (2019). Hands ahead in mind and motion: Active inference in peripersonal hand space. *Vision*, *3*(2). doi: 10.3390/vision3020015

Lohmann, J., & Butz, M. V. (2020). Hands in thought and motion. In *Proceedings for the 42nd annual meeting of the cognitive science society* (pp. 2822–2828).

Needham, T. (1993). A visual explanation of jensen's inequality. *The American Mathematical Monthly*, *100*(8), 768–771. doi: 10.1080/00029890.1993.11990484

Pérez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory* (pp. 1666–1670). doi: 10.1109/ISIT.2008.4595271

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extraclassical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. doi: 10.1038/4580

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Spence, C., Pavani, F., Maravita, A., & Holmes, N. (2004). Multisensory contributions to the 3-d representation of visuotactile peripersonal space in humans: Evidence from the crossmodal congruency task. *Journal of Physiology*, *98*(1), 171–189. doi: 10.1016/j.jphysparis.2004.03.008

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775. doi: 10.3758/BF03192967

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671. doi: 10.1080/09541440802205067

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. doi: 10.1037/0033-2909.133.2.273