1 Theoretical Framework

Our brain constantly receives incomplete and often ambiguous informations from sensory inputs, that are a noisy link between the brain and the external world.

While Helmholtz, in the late 19th century, was the first to propose a theory treating the brain as an inference machine [**?**], several works in the past few decades have further explored the idea of perception as an ongoing process of updating an inner generative model that infers the not accesible (hidden) causes of input stimuli [**???**]. One of the earliest concrete frameworks of this kind to describe the neural dynamics underlying perception in the visual cortex is the predictive coding model proposed by Rao and Ballard [**?**], which was later extended by Friston's free energy principle [**??**], which, in particular, takes action into account, thus creating a closed-loop scheme of perception.

This chapter provides a comprehensive introduction to the active inference framework, first introducing the work presented in [**?**], and then showing how Karl Friston placed it within the broader free energy principle, focusing in particular on the continuous time formulation (integrating in particular the excellent reviews [**???**]).

1.1 Rao and Ballard predictive coding model

A receptive field is defined [**?**] as a portion of sensory space that can elicit neuronal responses when stimulated. In the case of visual stimuli, receptive fields are two-dimensional region in the visual space that correlates with the activity of certain neurons. For example, cortical layers 2 and 3 in some mammalian visual cortexes, such as those of the monkeys [**?**], contain many neurons that respond optimally to line segments of a certain length (Fig. **??**). However, these neurons often exhibit reduced or eliminated responses when the same stimulus extends beyond their classical receptive fields. This 'extra-classical' effect generally occurs when stimulus properties at the center of the receptive field match those in the surrounding areas (Fig. **??**).

Using a specifically designed hierarchical neural network model [**?**], the authors show that such extra-classical receptive field effects could arise directly from the predictive coding structure of the network. The model proposes that neural networks learn the statistical regularities of the natural world and signal deviations from these regularities to higher processing centers, which reduce redundancy by removing the predictable, and thus redundant, components of the input signal.

1.2 The model

The following section describes the model using a different notation than in [**?**] in order to maintain consistency with the later sections.

Let us begin by considering as sensory input a grey-scale image with S pixels, denoted as a vector $\boldsymbol{s} = \{s\}_{i=1}^{S}$. Assuming that the cortex attempts to represent

the image in terms of hypothetical causes, represented by the K-dimensional vector $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^{K}$, we can express this relationship with the following equation:

$$\boldsymbol{s} = g(\boldsymbol{\mu}) + \mathcal{N}(\boldsymbol{s}; 0, \Sigma_s) \tag{1}$$

Here $g(\boldsymbol{\mu}) = U \cdot \boldsymbol{\mu}$ is the mapping function between sensory inputs and the state $\boldsymbol{\mu}$. In terms of a neural network, the coefficients μ_i correspond to the activities or firing rates of K neurons, whereas the columns of the linear operator U are the basis vectors for generating the images and correspond to the synaptic weights of the neurons. The second term of Eq. 1 is a noise term described by a Gaussian distribution with zero mean and Σ_s variance.

Because the dendritic arbors of neurons can only span a finite spatial extent, sensory inputs are divided into three overlapping Gaussian-windowed image patches[1], each of which is used as input (Fig. **??**) to three identical modules (Eq. 1) which comprise the first level of the network.

Above this level, the model includes a second one. It is assumed that it represents more abstract stimulus properties, represented by the activity of L neurons $\boldsymbol{\nu} = \{\nu_i\}_{i=1}^{L}$ related to the ones of the lower level through the following equation:

$$\boldsymbol{\mu} = f(\boldsymbol{\nu}) + \mathcal{N}(\boldsymbol{\mu}; 0, \Sigma_\nu) \tag{2}$$

The top-down prediction $f(\boldsymbol{\nu}) = U_\nu \cdot \boldsymbol{\nu}$ has a similar structure to g, and it is also accompanied by a noise term with a Gaussian distribution with zero mean and variance Σ_ν

The goal is to estimate, for each image presented, the coefficients $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and, on a longer time scale, learn appropriate basis vectors U and U_ν to describe efficiently a wide range of input images. Since the noise terms in Eq. 1 and 2 have been assumed Gaussian, it is possible to write the following optimization function:

$$E_l = \frac{1}{2}(\boldsymbol{s} - g(\boldsymbol{\mu}))^T \Sigma_s^{-1} (\boldsymbol{s} - g(\boldsymbol{\mu})) + \frac{1}{2}(\boldsymbol{\mu} - f(\boldsymbol{\nu}))^T \Sigma_\nu^{-1} (\boldsymbol{\mu} - f(\boldsymbol{\nu})) \tag{3}$$

It is important to note that this quantity is the negative logarithm of the probability of the data given the parameters of the model (negative log-likelihood) and that it is a weighted sum of quadratic forms of level-1 and level-2 prediction errors. Taking into account also Gaussian prior distributions for $\boldsymbol{\mu}$ and U, and using again their negative logarithms, it is possible to obtain the final optimization funciton

$$E = E_l + \frac{1}{2\Sigma_\mu} \sum_i \mu_i^2 + \frac{1}{2\Sigma_U} \gamma \sum_{i,j} U_{i,j}^2 \tag{4}$$

where Σ_μ and Σ_U are the variances of the respective Gaussian prior distributions. Note that, from a Bayesian perspective, maximizing E is equivalent to maximize the posterior probability of the model parameters given the input data.

[1] in the original work the authors set an horizontal offset of 5 pixels

An optimal estimate of $\boldsymbol{\mu}$ can be obtained through a classical gradient descent on E with respect to $\boldsymbol{\mu}$

$$\begin{aligned}\dot{\boldsymbol{\mu}} &= -k_\mu \frac{\partial E}{\partial \boldsymbol{\mu}} = \\ &= k_\mu \left[U^T \frac{(\boldsymbol{s} - g(\boldsymbol{\mu}))}{\Sigma_s} - \frac{(\boldsymbol{\mu} - f(\boldsymbol{\nu}))}{\Sigma_\nu} - \frac{\boldsymbol{\mu}}{\Sigma_\mu} \right]\end{aligned} \tag{5}$$

where k_μ is a learning rate. Therefore, in order to modify $\boldsymbol{\mu}$ toward the optimal estimate, the $(\boldsymbol{s} - g(\boldsymbol{\mu}))$ and the $(\boldsymbol{\mu} - f(\boldsymbol{\nu}))$ residual errors are needed.

Each level in this hierarchical model network thus, attempts to predict the responses at the next lower level via feedback connections (Fig. **??**). The error between this prediction and the actual response is then sent back to the higher level via feedforward connections. This error signal is used to correct the estimate of the input signal at each level. The prediction and error-correction cycles occur concurrently throughout the hierarchy, so that top-down information influences lower-level estimates, and bottom-up information influences higher-level estimates of the input signal. Lower levels operate on smaller spatial scales, whereas higher levels estimate signal properties at larger scales because a higher-level module predicts and estimates the responses of several lower-level modules (for example, three in Fig. **??**). Thus, the effective receptive field size of units increases progressively until the highest level, where the receptive field spans the entire input image. The underlying assumption here is that the external environment generates natural signals hierarchically via interacting hidden physical causes (object attributes such as shape, texture and luminance) at multiple spatial and temporal scales. The goal of a visual system then becomes the optimal estimation of these hidden causes at each scale for each input image and, on a longer time scale, the learning of the parameters governing the hierarchical generative model. In terms of the synaptic learning rule in particular, this correspond to the optimal estimate of the matrix U, that can be done agin by performing the gradient descent

$$\dot{U} = -k_U \frac{\partial E}{\partial U} = k_U \left[\frac{(\boldsymbol{s} - g(\boldsymbol{\mu}))}{\Sigma_s} \boldsymbol{\mu}^T - \frac{U}{\Sigma_U} \right] \tag{6}$$

with k_U that is, again, a learning rate.

The estimation of $\boldsymbol{\nu}$ and U_ν is analogous to what has just been done for $\boldsymbol{\mu}$ and U_ν, with the difference that the terms corresponding to the respective prior distributions miss. This is the case where the prior distribution has been set to Gaussian with arbitrary high variance.

Given the hierarchical organization of the visual cortex and the almost always reciprocal nature of cortico-cortical connections, the model just presented proposes the following hypothesis: feedback connections from a higher area (e.g., V2) to a lower area (e.g., V1) carry predictions of expected neural activity in V1, while feedforward connections transmit to V2 the residual activity in V1 that was not predicted by V2. In order to test this hypothesis, this three-level hierarchical network of predictive estimators was trained using image patches

extracted from five natural images (see Fig. **??**). The rationale behind this choice was that the response properties of visual neurons may be largely influenced by the statistical properties of natural images.

The training were perfermed by maximizing the posterior probability of generating the observed data: for any given input, the network converged to a set of neuronal responses that were optimal for predicting that specific input following the first-order differential equation of Eq. 5. These responses were subsequently utilized to adapt the synaptic basis vectors following Eq. 6. The same description applies to each level of the hierarchy, with each level predicting the inputs at its lower level using its set of learned basis vectors and, on a slower time scale, adapting these basis vectors to enable more accurate prediction of the inputs in the future.

1.3 Results

Overall, this approach allowed the network to learn a hierarchical internal model of its natural image inputs. Given a general input image, the initial predictions at any given level are based on an arbitrary random combination of the basis vectors, giving large error signals. To minimize this error, the network converges to the responses that best predict the current input by subtracting the prediction from the input (via inhibition) and propagating the residual error signal to the neurons at the next level, which integrate this error (like in Eq. 5) and generate a better prediction. In practice, the error signal in a biological system can be carried by error detecting neurons, which send feedforward connections from the lower level to the higher level. In the visual cortex, feedforward connections to a higher area generally arise from the superficial layers (such as layer 2/3). A relatively large number of neurons in layer 2/3 of striate cortex (V1) show endstopping and related extra-classical effects. To ascertain whether these observed neuronal responses can be functionally interpreted as residual error signals, responses of these level-1 error-detecting neurons where recorded when exposed to the image of a short bar lying within their receptive field (Fig. **??**). The solid box in the left panel represents the receptive field size of the level-1 neurons, whereas the dotted box represents the level-2 receptive field. The panels in Fig. **??** instead show the two components that determine the error signal. Many of the error-detecting neurons showed significant non-zero responses, demonstrating that feedback from level 2 could not completely predict the responses at level 1. On the other hand, when the bar stimulus extends beyond the classical receptive field into the flanking regions (right box in Fig. **??**), the same error-detecting neurons showed little or no response because the predictions from level 2 were much more accurate, with prediction errors close to zero. The reason lies in the fact that the network was trained on natural images, where short bars rarely occur in isolation. Rather, a bar in a small region of an image is usually part of a longer bar that extends into neighboring regions. Because the network was optimized for natural image statistics, the most accurate predictions are generated when the input's properties match those of these natural images. The continuation of the bar into the surrounding

region provides the necessary context for the bar in the center to be predicted, much as in the case of retinal center-surround prediction mechanisms. Without this contextual information in the surrounding region, the higher level cannot accurately predict the bar in the center. The short bar thus elicits a relatively large response from the error-detecting neurons as compared to the longer bar.

1.4 The free energy principle

A significant breakthrough in the field of predictive coding theory occurred when it was realized that the predictive coding algorithm could be framed as an approximation of Bayesian inference, utilizing Gaussian generative models [**?**]. Friston's approach, importantly, redefines the predominantly heuristic model proposed by Rao and Ballard using the framework of variational Bayesian inference. This not only allows for a comprehensive theoretical understanding of the algorithm but it also strengthen its connection to the broader concept of the Bayesian Brain [**??**]. Friston demonstrated that the energy function in Rao and Ballard's model can be interpreted as a variational free-energy, which is minimized through variational inference. This connection establishes that predictive coding directly engages in approximate Bayesian inference to deduce the underlying causes of sensory signals, thereby offering a mathematically precise characterization of the Helmholtzian notion of perception as inference.

To formalize this let us assume that the goal of an agent is to determine the probability distribution of a latent state $\boldsymbol{x}$ given some sensory inputs $\boldsymbol{s}$:

$$p(\boldsymbol{x}|\boldsymbol{s}) = \frac{p(\boldsymbol{x},\boldsymbol{s})}{p(\boldsymbol{s})} = \frac{p(\boldsymbol{s}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{s})} \tag{7}$$

with

- $P(\boldsymbol{x},\boldsymbol{s})$ *joint density*, beliefs about the states assumed to be encoded by the agent. In the framework it is referred to as *generative model*.
- $p(\boldsymbol{x}|\boldsymbol{s})$ *Posterior*, i.e. probability distribution of hidden causes x given observed sensory data;
- $p(\boldsymbol{s}|\boldsymbol{x})$ *Likelihood*, i.e. organism's assumptions about sensory input $\boldsymbol{s}$ given the hidden causes $\boldsymbol{x}$;
- $p(\boldsymbol{x})$ *Prior*, i.e. agent's beliefs about hidden causes before that $\boldsymbol{s}$ are received;
- $p(\boldsymbol{s}) = \int p(\boldsymbol{s}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ *marginal likelihood* (normalization factor). It is also referred to as *evidence* or *model evidence*, since it effectively scores the likelihood of the data under a given model averaged over all possible values of the model parameters.

The first problem with this exact Bayesian scheme is that calculating $p(\boldsymbol{s})$ is often impossible. However, since the agent does not necessarily need to compute

the complete posterior distribution, but only needs to find the hidden state, its goal is to find:

$$\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{s}) \tag{8}$$

However, this goal may also present some challenges. In fact, the distribution may not have a standard shape or have summary statistics.

Another biologically plausible technique to approximate the posterior $p(\boldsymbol{x}|\boldsymbol{s})$ consist in using an auxiliary distribution $q(\boldsymbol{x})$, called *recognition density*, that has to be optimized to became a good approximation of the posterior.

In order to do this the Kullback-Leibler divergence is minimized:

$$\begin{aligned} D_{KL}(\, q(\boldsymbol{x}) \,||\, p(\boldsymbol{x}|\boldsymbol{s}) \,) &= \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{s})} d\boldsymbol{x} \\ &= \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})p(\boldsymbol{s})}{p(\boldsymbol{x},\boldsymbol{s})} d\boldsymbol{x} \\ &= \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x},\boldsymbol{s})} d\boldsymbol{x} + \ln p(\boldsymbol{s}) \int q(\boldsymbol{x}) d\boldsymbol{x} \\ &= \mathcal{F} + \ln p(\boldsymbol{s}) \end{aligned} \tag{9}$$

where

- the term

$$\mathcal{F} \equiv \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x},\boldsymbol{s})} d\boldsymbol{x} = \langle \ln q(\boldsymbol{x}) \rangle_q - \langle \ln p(\boldsymbol{x},\boldsymbol{s}) \rangle_q \tag{10}$$

  is the *variational free energy*[2], a quantity that depends on the recognition density and the knowledge about the environment i.e. the joint density $p(\boldsymbol{s},\boldsymbol{x}) = p(\boldsymbol{s}|\boldsymbol{x})p(\boldsymbol{x})$ that for now we are assuming the agent has.

- $\ln p(\boldsymbol{s})$ is the *log-evidence*, a term independent with respect to the recognition density $q(\boldsymbol{x})$. Thus, with fixed sensory inputs, minimizing $\mathcal{F}$ with respect to $q(\boldsymbol{x})$ will in turn minimize the $D_{KL}(\, q(\boldsymbol{x}) \,||\, p(\boldsymbol{x}|\boldsymbol{s}) \,)$ (in general minimizing $\mathcal{F}$ results in a reduction of $D_{KL}(\, q(\boldsymbol{x}) \,||\, p(\boldsymbol{x}|\boldsymbol{s}) \,)$).

To have a better intuition about $\mathcal{F}$, it is possible to re-write Eq. 10 in the following ways

$$\begin{aligned} \mathcal{F} &\equiv \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x},\boldsymbol{s})} d\boldsymbol{x} = \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{s}|\boldsymbol{x})p(\boldsymbol{x})} d\boldsymbol{x} \\ &= \underbrace{D_{KL}(\, q(\boldsymbol{x}) || p(\boldsymbol{x}) \,)}_{complexity} - \underbrace{\langle \ln p(\boldsymbol{s}|\boldsymbol{x}) \rangle_q}_{accuracy} \\ &= \underbrace{D_{KL}(\, q(\boldsymbol{x}) || p(\boldsymbol{x}|\boldsymbol{s}) \,)}_{divergence} - \underbrace{\ln p(\boldsymbol{s})}_{evidence} \end{aligned} \tag{11}$$

[2]In machine learning the negative of this quantity is called evidence lower bound (ELBO), and is maximized instead.

Each of these formulations of variational free energy offers useful intuitions about what free energy minimization means.

The second line of Eq. 11 emphasizes the interpretation of free energy minimization as finding the best explanation for sensory data, which must be the simplest (i.e. minimally complex) explanation that is able to accurately account for the data (Occam's razor): a negative value of a term called *accuracy* is present, that contains a likelihood that has to be maximized, with a Kullback-Leibler divergence term called *complexity* which penalizes deviations from the Bayesian prior. In many machine learning algorithms this decomposition is often utilized and explicitly optimized [**?**].

The final line instead expresses the free energy as a bound on negative log *evidence*. The free energy in fact is an upper bound of this quantity, where the bound is the divergence between $q(\boldsymbol{x})$ and the posterior probability $p(\boldsymbol{x}|\boldsymbol{s})$. This offers a formal motivation for perceptual inference as one way to lower free energy by optimizing our approximate posterior $q(\boldsymbol{x})$ as much as possible. Moreover, as we will see in Sec. 1.8, perceptual inference is not the only way to minimize free energy. It can be also possible to change the log evidence term through acting to change sensory data. This decomposition is interesting from a cognitive perspective, since minimizing divergence and maximizing evidence map to the two complementary sub-objectives of perception and action, respectively.

1.5 Predictive coding as variational inference

For the sake of clarity, we will build the framework in the one dimensional case, explaining in subsequent chapters, when necessary, the corresponding variables and functions in higer dimensions.

1.5.1 Laplace approximation

Often optimizing $\mathcal{F}$ for arbitrary $q(x)$ is particularly complex. Moreover, it is assumed that neural activity encode a parametrised model – with finite numbers of parameters. For these reasons, a common approximation is to assume that the recognition density take a Gaussian form.

Assuming that $q(x)$ has a peak at point μ, the Taylor-expansion of the logarithm around this peak is

$$\ln q(x) \simeq \ln q(\mu) - \frac{1}{2}\frac{(x-\mu)^2}{\Sigma} \tag{12}$$

with

$$\frac{1}{\Sigma} = -\frac{\partial^2}{\partial x^2} \ln q(x)\bigg|_{x=\mu} \tag{13}$$

Now it is possible to approximate the probability distribution $q(x)$ with the distribution

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi\Sigma}}\, e^{\frac{(x-\mu)^2}{2\Sigma}} \tag{14}$$

i.e. a Gaussian distribution that has been normalized using the factor $q(\mu)\sqrt{2\pi\Sigma}$. Now the variational free energy, starting from Eq. 10, can be written as follows

$$\begin{aligned}\mathcal{F} &= \langle \ln q(\boldsymbol{x}) \rangle_q - \langle \ln p(\boldsymbol{x}, \boldsymbol{s}) \rangle_q \\ &\approx \int \mathcal{N}(x; \mu, \Sigma)(-\frac{1}{2}\ln(2\pi\Sigma) - \frac{(x-\mu)^2}{2\Sigma})dx - \int \mathcal{N}(x; \mu, \Sigma) \ln p(x, s) dx \\ &= -\frac{1}{2}\ln(2\pi\Sigma) - \frac{1}{2\Sigma}\int \mathcal{N}(x; \mu, \Sigma)(x-\mu)^2 dx - \int \mathcal{N}(x; \mu, \Sigma) \ln p(x, s) dx \\ &= -\frac{1}{2}\ln(2\pi\Sigma) - \frac{1}{2} - \int \mathcal{N}(x; \mu, \Sigma) \ln p(x, s) dx\end{aligned} \tag{15}$$

To obtain an analytical model, additional simplifications and assumptions are required in order to evaluate the final term[3].

Assuming that $p(x, s)$ is a smooth function of x, it is possible to consider the integrated function appreciably non-zero only near the peak of the q. So, using a second order Taylor expansion of the $L(x, s) \equiv -\ln p(x, s)$ around $x = \mu$ we ca write

$$L(x, s) \approx L(\mu, s) + \left[\frac{dL(x, s)}{dx}\right]_{x=\mu}(x-\mu) + \frac{1}{2}\left[\frac{d^2L(x, s)}{dx^2}\right]_{x=\mu}(x-\mu)^2 \tag{16}$$

This implies that

$$\begin{aligned}-\int \mathcal{N}(x; \mu, \Sigma) \ln p(x, s) dx \approx \int \mathcal{N}(x; \mu, \Sigma)\Big\{ & L(\mu, s) + \left[\frac{dL(x, s)}{dx}\right]_{x=\mu}(x-\mu) + \\ & + \frac{1}{2}\left[\frac{d^2L(x, s)}{dx^2}\right]_{x=\mu}(x-\mu)^2\Big\}dx \\ = L(\mu, s) &+ \left[\frac{dL(x, s)}{dx}\right]_{x=\mu}\left(\int \mathcal{N}(x; \mu, \Sigma) x\, dx - \mu\right) + \\ &+ \frac{1}{2}\left[\frac{d^2L(x, s)}{dx^2}\right]_{x=\mu}\int \mathcal{N}(x; \mu, \Sigma)(x-\mu)^2\, dx \\ = L(\mu, s) &+ \frac{1}{2}\left[\frac{d^2L(x, s)}{dx^2}\right]_{x=\mu}\Sigma\end{aligned} \tag{17}$$

Now is possible to rewrite the variational free energy as

$$\mathcal{F}(\mu, \Sigma, s) \approx L(\mu, s) + \frac{1}{2}\left(\left[\frac{d^2L(x, s)}{dx^2}\right]_{x=\mu}\Sigma - \ln(2\pi\Sigma) - 1\right) \tag{18}$$

with $L(\mu, s)$ said *Laplace-encoded energy*, and the variational free energy written as a function and not anymore as a functional.

[3]Later we will discuss the implications of this issue for the interpretation of brain functions.

Since the goal is to minimize the Kullback-Leibler divergence trough the minimization of the variational free energy, is possible to simplify further removing the Σ dependency taking the derivative with respect this and imposing $\frac{d\mathcal{F}}{d\Sigma} = 0$

$$\frac{d\mathcal{F}}{d\Sigma} = \frac{1}{2}\left(\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu} - \frac{1}{\Sigma}\right) = 0 \tag{19}$$

$$\Rightarrow \Sigma = \left[\frac{d^2L(x,s)}{dx^2}\right]^{-1}_{x=\mu} \equiv \Sigma^* \tag{20}$$

The final form of the variational free energy is then

$$\mathcal{F} \approx L(\mu, s) - \frac{1}{2}\ln\left(2\pi\Sigma^*\right) , \tag{21}$$

that can be further simplified if the L function has a second order polynomial form[4], that implies that the second-order derivative of L with respect to x results in a constant factor that is useless and can be ignored[5], leading to

$$\mathcal{F} \approx L(\mu, s) \tag{22}$$

Simple static model with entropy term Starting from a joint density with the following form

$$p(x,s) = p(s|x)p(x) = \mathcal{N}(s; f_0(x), \Sigma_s)\mathcal{N}(x; \nu, \Sigma_x) \tag{23}$$

it is possible to write the Laplace-encoded energy

$$\begin{aligned} L(\mu, s) &= -\ln p(s|\mu) - \ln p(\mu) \\ &= \frac{1}{2}\ln(2\pi\Sigma_s) + \frac{(s-g(\mu))^2}{2\Sigma_s} + \frac{1}{2}\ln(2\pi\Sigma_x) + \frac{(\mu-\nu)^2}{2\Sigma_x} \\ &= \frac{\varepsilon_s^2}{2\Sigma_s} + \frac{\varepsilon_\mu^2}{2\Sigma_x} + \frac{1}{2}\ln\left(\Sigma_s\Sigma_x\right) + \ln(2\pi) , \end{aligned} \tag{24}$$

and consequently the variational free energy

$$\mathcal{F} \approx \frac{1}{2}\left[\frac{\varepsilon_s^2}{\Sigma_s} + \frac{\varepsilon_\mu^2}{\Sigma_x} + \ln\left(\Sigma_s\Sigma_x\right) - \ln(2\pi\Sigma^*)\right] . \tag{25}$$

where in this case Σ^* is equal to

$$\begin{aligned} \Sigma^* \equiv \left[\frac{d^2L(x,s)}{dx^2}\right]^{-1}_{x=\mu} &= \frac{d}{d\mu}\left[\frac{\varepsilon_s}{\Sigma_s}\frac{d\varepsilon_s}{d\mu} + \frac{\varepsilon_\mu}{\Sigma_x}\frac{d\varepsilon_\mu}{d\mu}\right] \\ &= \frac{d}{d\mu}\left[-\frac{\varepsilon_s}{\Sigma_s}\frac{dg(\mu)}{d\mu} + \frac{\varepsilon_\mu}{\Sigma_x}\right] \\ &= \frac{1}{\Sigma_s}\left(\frac{dg(\mu)}{d\mu}\right)^2 - \frac{\varepsilon_s}{\Sigma_s}\frac{d^2g(\mu)}{d\mu^2} + \frac{1}{\Sigma_x} \end{aligned} \tag{26}$$

[4] as we will see in Sec.(1.6)

[5] let us remind that the final goal is to minimize $\mathcal{F}$ with respect to x

Now this term can be neglected if, and only if, $dg(\mu)/d\mu = 0$ and you are not modifying $p(x)$ precisions.

About Laplace Approximation Another approximation that can be done, essentially identical to the Laplace approximation considering that it arrives at the same expression for the variational free energy, is the one that consider the recognition density equal to dirac-delta distribution

$$q(x) \approx \delta(x - \mu) \tag{27}$$

The main difference between these approximations is the value of the resulting entropy term of Eq. 10, which under the dirac-delta assumion is a null term, while under the Laplace approximation is nonzero but constant with respect to parameters being optimized. Both this procedures lead to Eq.(22), that would imply that that the brain represents the hidden state only through the most likely cause, and nothing else about the recognition distribution. However, as we are going to see, the uncertainties will be encoded directly in the form of the joint density.

1.6 Building the generative model and variational free energy minimization

Thanks to the Laplace approximation, we have been able to write the variational free energy in terms of the Laplace-encoded energy $L(\mu, s)$, that in turn depends on the joint density $p(x, s)$. In this function are encoded brain beliefs about the environmental causes of the sensory input and the beliefs *a priori* about environmental states.

Therefore, what needs to be built is a *generative model*, that is a model in which is encoded how the brain believe the world works and where all the hypothesis about the agent's behaviour are formalized.

1.6.1 Static Model

Let us consider a simple case of an agent that, as generative model, represents the hidden state of the environment through the variable x, which in turn cause the data received by a sensory channel s. As we have seen in Sec.(1.5.1), the brain will represent the environment only through the inner state μ, and what remains to do is to explicit the mapping between brain states and sensory data that will allows to make explicit the joint density.

Assuming that the agent believes that its sensory input are generated by

$$s = g(x) + \mathcal{N}(s; 0, \Sigma_s)\,, \tag{28}$$

with g being a generic function that expresses the relation between states and sensory input, to which is summed a noise represented by the normal distribution with zero mean and variance Σ_s. This assumption means that we can write

$$p(s|x) = \mathcal{N}(s; g(x), \Sigma_s) = \frac{1}{\sqrt{2\pi\Sigma_s}}\, e^{-\frac{(s-g(x))^2}{2\Sigma_s}}\,. \tag{29}$$

Moreover let us also assume that the agent also has a prior knowledge regarding the environmental state given by ν that is linked with the inner state through

$$x = \nu + \mathcal{N}(x; 0, \Sigma_x) \tag{30}$$

$$\Rightarrow p(x) = \mathcal{N}(x; \nu, \Sigma_x) = \frac{1}{\sqrt{2\pi\Sigma_x}} e^{-\frac{(x-\nu)^2}{2\Sigma_x}} . \tag{31}$$

Now that we have specified a likelihood and a prior, is possible to determine the joint density

$$p(x, s) = p(s|x)p(x) \tag{32}$$

and consequently the Laplace-encoded energy

$$\begin{aligned} L(\mu, s) &= -\ln p(s|\mu) - \ln p(\mu) \\ &= \frac{1}{2}\ln(2\pi\Sigma_s) + \frac{(s - g(\mu))^2}{2\Sigma_s} + \frac{1}{2}\ln(2\pi\Sigma_x) + \frac{(\mu - \nu)^2}{2\Sigma_x} \\ &= \frac{\varepsilon_s^2}{2\Sigma_s} + \frac{\varepsilon_\mu^2}{2\Sigma_x} + \frac{1}{2}\ln\left(\Sigma_s\Sigma_x\right) + \ln(2\pi)\,, \end{aligned} \tag{33}$$

where the ε terms are said *prediction errors* and measure the discrepancy respectively between the actual sensory data s and the outcome of its prediction $g(x)|_{x=\mu}$ and between μ itself and its prior expectation ν. Therefore the former ε_s describes sensory prediction errors, the latter ε_μ model prediction errors (i.e. how brain states deviate from their expectation) and each one is weighted with the the corresponding inverse of the variance Σ_s^{-1} and Σ_x^{-1} (which are often said *precisions*).
As said at the end of Sec.(1.5.1), since the L function has a quadratic form, it is possible to ignore all the terms apart from the following

$$\mathcal{F} \approx \frac{1}{2}\left[\frac{\varepsilon_s^2}{\Sigma_s} + \frac{\varepsilon_\mu^2}{\Sigma_\mu} + \ln\left(\Sigma_s\Sigma_\mu\right)\right] . \tag{34}$$

The last thing that remains to do is to find a biologically plausible mechanism to minimize variational free energy.
In the free energy principle framework, it is proposed that the innate dynamics of the neural activity evolves in such a way that it implements a gradient descent scheme on the variational free energy.
In particular, in the static model case, a brain state μ is updated between two (internal) sequential steps t and $t + dt$ as

$$\dot{\mu} = -k_\mu \frac{\partial \mathcal{F}}{\partial \mu} = k_\mu \left[\frac{\varepsilon_s}{\Sigma_s}\frac{\partial g(\mu)}{\partial \mu} - \frac{\varepsilon_\mu}{\Sigma_\mu}\right], \tag{35}$$

with k learning rate parameter that has to be tuned and $\frac{\partial \mathcal{F}}{\partial \mu}$ goes to zero when a minimum of the $\mathcal{F}$ function is reached.

In many practical cases, it is possible to relax the assumption that the agent has a fixed generative model (Eq. 32), and let it learn in parallel while inferring

μ. Writing the g function with an explicit parameter $g(x;\theta_g)$ and expressing the state x not only as a prior given by ν but a more generic $f(\nu;\theta_f)$, it is possible to write the update rules also of the model parameters θ_g and θ_f

$$\dot{\theta}_g = -k_g \frac{\partial \mathcal{F}}{\partial \theta_g} = k_g \frac{\varepsilon_g}{\Sigma_s} \frac{\partial g(\mu;\theta_g)}{\partial \theta_g} \tag{36}$$

$$\dot{\theta}_f = -k_f \frac{\partial \mathcal{F}}{\partial \theta_f} = k_f \frac{\varepsilon_f}{\Sigma_x} \frac{\partial f(\nu;\theta_f)}{\partial \theta_f} \tag{37}$$

For example, in the [**?**] model the g function was parametrised by the matrix U (Eq. 1), while the f function was parametrized by the matrix U_ν (Eq. 2), and the only difference from the current formulation is the presence in Eq. 4 of the priors on $\boldsymbol{\mu}$ and U. Writing the following generative model (i.e. joint probability density form) as follows,

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{s}) &= p(\boldsymbol{s}|\boldsymbol{x})p(\boldsymbol{x})p(\boldsymbol{\theta}_g) \\ &= \mathcal{N}(\boldsymbol{s};U\cdot\boldsymbol{x},\Sigma_s)\mathcal{N}(\boldsymbol{x};U_\nu\cdot\boldsymbol{\nu},\Sigma_\nu)\mathcal{N}(\boldsymbol{x};0,\Sigma_x)\mathcal{N}(U;0,\Sigma_U) \end{aligned} \tag{38}$$

leads to a Laplace-encoded energy with the same form of the optimization function of Eq. 4.

While it is possible to update μ and θ simultaneously, as done in [**?**], it is often better to treat predictive coding as an EM algorithm [**?**], optimizing μ, with fixed θ_g and θ_f until close to convergence, and then run the updates on the parameters with fixed μ for a short while. This implicitly enforces a separation of timescales upon the model where μ is seen as a dynamical variable which change quickly while the θ_g and θ_f are slowly changing parameters. It correspond on a mean-field approximation where the set of slowly changing parameters are treated as conditionally independent with respect to hidden states and sensory inputs.

1.6.2 Dynamic Model

Biological agents rarely deal with stationary conditions, which is why it is important to be able to have generative models that can deal with dynamic environments: let us now formulate a possible implementation of inference in a dynamically changing environment.

In order to describe a dynamical system, it is necessary to have at least two environmental variables: x and its first-order derivative with respect to time, $\frac{dx}{dt} = x'$. Therefore, the agent must possess prior knowledge of the environmental state x and also model its dynamics.

In a dynamic case, the variable x is typically allowed to vary without any boundary. Therefore, a flat prior, indicating zero prior knowledge, is usually associated with it and the dynamic of x can be described using a Langevin-type equation:

$$\frac{dx}{dt} = f(x) + w_x \,. \tag{39}$$

Usually, in active inference literature, agent's internal representation of the various order of motion are indicated using the superscript, while how these variables are effectively updated during the free energy minimization process is indicated using the dot notation (for example, as we will see, the representation of the first order of an internal variable μ will be expressed with the term μ', while the temporal increment during the inference process of μ will be denoted with $\dot{\mu}$ and the temporal increment of μ' with $\dot{\mu}'$).

As in the static case, is assumed that the agent believes its sensory input are generated in a similar manner with respect to Eq.(28), in particular

$$s(x) = g(x) + w_s \,. \tag{40}$$

Both w_s and w_x are again terms representing Gaussian noise with zero mean and variances respectively equal to Σ_s and Σ_x.

The joint probability can then be written

$$p(x, x', s) = p(s|x)p(x'|x)p(x) = C \cdot \mathcal{N}(s; g(x), \Sigma_s)\mathcal{N}(x'; f(x), \Sigma_{\dot{x}}) \tag{41}$$

leading to a Laplace-encoded energy

$$\begin{aligned} L(\mu', \mu, s) &= \frac{(s - g(\mu))^2}{2\Sigma_s} + \frac{1}{2}\ln(2\pi\Sigma_s) + \frac{(\mu' - f(\mu))^2}{2\Sigma_\mu} + \frac{1}{2}\ln(2\pi\Sigma_\mu) + C \\ &= \frac{\varepsilon_s^2}{2\Sigma_s} + \frac{\varepsilon_\mu^2}{2\Sigma_\mu} + \frac{1}{2}\ln(\Sigma_s\Sigma_\mu) + \ln(2\pi) + C \end{aligned} \tag{42}$$

and consequently an approximated variational free energy

$$\mathcal{F} \approx \frac{\varepsilon_s^2}{2\Sigma_s} + \frac{\varepsilon_\mu^2}{2\Sigma_\mu} + \frac{1}{2}\ln(\Sigma_s\Sigma_\mu) \tag{43}$$

Here we have modified the notation for ε_x and Σ_x to ε_μ and Σ_μ, respectively. This change was made because in generative models, one only needs to work with the expected values of the corresponding variables, which are never explicitly shown.

1.6.3 Generalised state-space model

To effectively represent complex dynamics of the generative process, it is possible to improve the agent's model by using generalized coordinate of motion **[??]** beyond the first order. For example, if the brain represents beliefs about the position of an object, a generalized coordinates model would also include beliefs about its velocity, acceleration, jerk, and so on. However, this approach requires a departure from considering the dynamics of the system as a single stochastic equation with white noise.

Therefore, by representing the state of the dynamical system in terms of increasingly higher order derivatives of its state variables and employing a local

linearity approximation on higher orders of motion[6] to suppress non-linear terms in the partial derivatives, it becomes possible to obtain

$$\begin{array}{ll} s = g(x) + w_s(t) & x' = f(x) + w_x(t) \\ s' = \dfrac{\partial g(x)}{\partial x} x' + w'_s & x'' = \dfrac{\partial f(x)}{\partial x} x' + w'_x \\ s'' \simeq \dfrac{\partial g(x)}{\partial x} x'' + w''_s & x''' \simeq \dfrac{\partial f(x)}{\partial x} x'' + w''_x \\ \vdots & \vdots \end{array} \tag{44}$$

where we have used the notation

$$s' = \frac{ds}{dt}\,, \qquad x' = \frac{dx}{dt}\,, \qquad s'' = \frac{d^2 s}{dt^2}\,, \qquad x'' = \frac{d^2 x}{dt^2}\,, \quad \ldots \tag{45}$$

and where $w_s, w'_s, w''_s, \ldots, w_x, w'_x, w''_x, \ldots$ are the noise sources at each dynamic order. Considering the previous linear approximation as equalities, Eq.(44) can be expressed in the more compact form

$$\tilde{s} = \tilde{g}(\tilde{x}) + \tilde{w_s} \quad , \quad D\tilde{x} = \tilde{f}(\tilde{x}) + \tilde{w_x} \tag{46}$$

using the notation

$$\begin{array}{l} \tilde{s} = (s, s', s'', \ldots) \\ \tilde{x} = (x, x', x'', \ldots) \\ D\tilde{x} = (x', x'', x''', \ldots) \\ \tilde{g}(\tilde{x}) = (g(x), \dfrac{\partial g(x)}{\partial x} x', \dfrac{\partial g(x)}{\partial x} x'', \ldots \\ \tilde{f}(\tilde{x}) = (f(x), \dfrac{\partial f(x)}{\partial x} x', \dfrac{\partial f(x)}{\partial x} x'', \ldots \end{array} \tag{47}$$

Considerations about noise terms As seen in Appendix **??**, a differentiable Gaussian stochastic process W with zero mean, Σ_w variance and auto-correlation function $\rho(h) \equiv < W(t+h)W(T) >_s$, can be written as a multivariate Gaussian distribution with covariance matrix (Eq. **??**)

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma & 0 & \rho''(0) & \ldots \\ 0 & -\rho''(0) & 0 & \\ \rho''(0) & 0 & \rho''''(0) & \\ \vdots & & & \ddots \end{bmatrix} \tag{48}$$

with $\rho''(0)$ second derivative of the auto-correlation function of the fluctuations, evaluated at zero, that is a ubiquitous measure of roughness in the theory of

[6] Without this approximation the model would scale-up very quickly becoming complicated and unwieldy fairly quickly. This approximation becomes exact when f and g are linear.

stochastic processes [**?**]. Note that when noise terms at different order are uncorrelated, the curvature (and higher derivatives) of the auto-correlation becomes large (i.e. $\rho''(0) \to \infty$). In this instance, the variances of the temporal derivatives fall to zero and the variational energy is determined by, and only by, the magnitude of the prediction errors on the causes and the first-order motion of the hidden states. This limiting case is assumed by conventional state-space models used in Bayesian filtering; it corresponds to the assumption that the fluctuations are independent[7]. Although, this is a convenient assumption for conventional schemes and appropriate for physical systems with Brownian processes, it is less plausible for biological and other systems, where random fluctuations are themselves the product of other dynamical systems. For convenience then, in [**?**] it is assumed that the noise correlation is due to a Gaussian filter

$$S(\gamma) = \begin{bmatrix} 1 & 0 & -\frac{1}{2}\gamma & \dots \\ 0 & \frac{1}{2}\gamma & 0 & \\ -\frac{1}{2}\gamma & 0 & \frac{3}{4}\gamma^2 & \\ \vdots & & & \ddots \end{bmatrix} \tag{49}$$

where γ is the precision (inverse variance) parameter of the filter, which increases with roughness. Assuming zero cross-correlation between $\tilde{w}_s$ and $\tilde{w}_\mu$, the final covariance matrices of the noise terms will be then

$$\tilde{\Sigma}_s = S(\gamma)\Sigma_s \quad \tilde{\Sigma}_\mu = S(\gamma)\Sigma_\mu \tag{50}$$

Typically, $\gamma > 1$, which ensures the variances of higher-order derivatives to diverge quickly. This is important because it enables us to truncate the representation in generalised coordinates to a relatively low order (6 orders are sufficient for most systems, but truncating to second-order generally doesn't lead to an accuracy loss)

The variational free energy in the generalised state-space model, will then be approximated by

$$\mathcal{F} = \frac{1}{2}\tilde{\varepsilon}_s^T \tilde{\Pi}_s \tilde{\varepsilon}_s + \frac{1}{2}\tilde{\varepsilon}_\mu^T \tilde{\Pi}_\mu \tilde{\varepsilon}_\mu + \frac{1}{2}\ln(\det \tilde{\Sigma}_s \det \tilde{\Sigma}_\mu) \tag{51}$$

with $\tilde{\varepsilon}_s = \tilde{s} - \tilde{g}(\tilde{\mu})$, $\tilde{\varepsilon}_\mu = D\tilde{\mu} - \tilde{f}(\tilde{\mu})$, $\tilde{\Pi}_s = \tilde{\Sigma}_s^{-1}$ and $\tilde{\Pi}_\mu = \tilde{\Sigma}_\mu^{-1}$.

Now, last thing remaining is the minimization process. It has been shown [**?**] that the optimal (equilibrium) solution can be reached through the (modified) gradient descend

$$\tilde{\mu}^{t+\Delta t} = \tilde{\mu}^t + D\tilde{\mu} - k_\mu \frac{\partial \mathcal{F}}{\partial \tilde{\mu}} \tag{52}$$

This adjustment can be understood considering that, since we are minimizing the components of a generalised state representing a trajectory rather than a static state, variables are in a moving framework of reference, and the minimization is achieved for $\dot{\tilde{\mu}} = D\tilde{\mu}$, rather than for $\dot{\tilde{\mu}} = 0$ (condition required in standard state-space formulations).

[7] This correspond to a purely diffusion process, like the Wiener process. This case is also called white noise scenario

1.7 Multi-layer models

The concepts discussed thus far can be further expanded upon by incorporating a hierarchical structure into the generative model. As demonstrated in the basic model of [**?**] (Sec. 1.1), even a single layer was able to expand receptive fields. By adding more layers, drawing inspiration from the cortical hierarchical organization, it becomes feasible to encode complex abstractions and effectively manage inherently hierarchical dynamics, similar to how humans naturally perceive them.

Regarding the implementation of an upper layer, the same reasoning can be applied as for the development of the generative model discussed thus far. One can begin with a static model, which involves a specific probability distribution on the internal variables of the model. Alternatively, one can develop a second dynamical model that aims to infer the state of the preceding layer.

Indicating with $x^{(1)}$ the first layer hidden state, v the *hidden cause*[8], and with $x^{(2)}$ the hidden state of the second layer, we can rewrite the generative model as follow

$$\begin{aligned} s(x^{(1)}, v) &= g^{(1)}(x^{(1)}, v) + w_s \\ \frac{dx^{(1)}}{dt} &= f^{(1)}(x^{(1)}, v) + w_{x^{(1)}} \\ v(x^{(2)}) &= g^{(2)}(x^{(2)}) + w_v \\ \frac{dx^{(2)}}{dt} &= f^{(2)}(x^{(2)}) + w_{x^{(2)}} \end{aligned} \tag{53}$$

leading to the approximated variational free energy (Here again we have substituted for the various x and v the corresponding mean values μ and ν in the notation)

$$\begin{aligned} \mathcal{F} =& \frac{1}{2}\varepsilon_s^T \Pi_s \varepsilon_s + \frac{1}{2}\varepsilon_{\mu^{(1)}}^T \Pi_{\mu^{(1)}} \varepsilon_{\mu^{(1)}} + \frac{1}{2}\varepsilon_\nu^T \Pi_\nu \varepsilon_\nu + \frac{1}{2}\varepsilon_{\mu^{(2)}}^T \Pi_{\mu^{(2)}} \varepsilon_{\mu^{(2)}} \\ &+ \frac{1}{2}\ln(\det \Sigma_s \det \Sigma_{\mu^{(1)}} \det \Sigma_\nu \det \Sigma_{\mu^{(2)}}) \end{aligned} \tag{54}$$

and to the minimization process

$$\begin{aligned} \dot{\mu}^{(1)} &= \mu'^{(1)} - k_{\mu^{(1)}} \left[-\frac{\varepsilon_s}{\Sigma_s}\frac{\partial g^{(1)}}{\partial \mu^{(1)}} - \frac{\varepsilon_{\mu^{(1)}}}{\Sigma_{\mu^{(1)}}}\frac{\partial f^{(1)}}{\partial \mu^{(1)}} \right] \\ \dot{\mu'}^{(1)} &= -k_{\mu'^{(1)}} \frac{\varepsilon_{\mu^{(1)}}}{\Sigma_{\mu^{(1)}}} \\ \dot{\nu} &= -k_\nu \left[-\frac{\varepsilon_s}{\Sigma_s}\frac{\partial g^{(1)}}{\partial \nu} - \frac{\varepsilon_{\mu^{(1)}}}{\Sigma_{\mu^{(1)}}}\frac{\partial f^{(1)}}{\partial \nu} + \frac{\varepsilon_n u}{\Sigma_n u} \right] \\ \dot{\mu}^{(2)} &= \mu'^{(2)} - k_{\mu^{(2)}} \left[-\frac{\varepsilon_\nu}{\Sigma_\nu}\frac{\partial g^{(2)}}{\partial \mu^{(2)}} - \frac{\varepsilon_{\mu^{(2)}}}{\Sigma_{\mu^{(1)}}}\frac{\partial f^{(2)}}{\partial \mu^{(2)}} \right] \\ \dot{\mu'}^{(2)} &= -k_{\mu'^{(2)}} \frac{\varepsilon_{\mu^{(2)}}}{\Sigma_{\mu^{(2)}}} \end{aligned} \tag{55}$$

[8] It has the same role of model parameters of Eq. 36 and 37

We see that the dynamics for the variational means μ depend only on the prediction errors at their layer and the prediction errors on the level below. Intuitively, we can think of the μ as trying to find a compromise between causing error by deviating from the prediction from the layer above, and adjusting their own prediction to resolve error at the layer below. In a neurally-implemented hierarchical predictive coding network, prediction errors would be the only information transmitted upwards from sensory data towards latent representations, while predictions would be transmitted downwards.

let us conclude the section observing that these models have the potential for further expansion by incorporating generalized state-space models into multiple layers. Additionally, employing a mean-field approximation with slower time scales could enable the model to learn the different covariance matrices.

1.8 Action

So far, we have considered what happens when we perform inference, hence selecting the model on the basis of its capacity to minimize surprise (or equivalently, maximize model evidence of Eq. 11). However, surprise does not only depend on the model, but it also depends on the data. By acting on the world to change the way in which data are generated, we can ensure a model is fit for purpose by choosing those data that are least surprising under our model. In particular, in active inference, action is described as a problem of optimal control that essentially mirrors perception by changing observations s to better match expected hidden states μ. This process is based on the general assumption that, from the perspective of an agent, observations s are affected by actions a (s is a function of $a \Rightarrow s(a)$).

Action is then performed with the objective of consistently minimizing the same cost function, namely the variational free energy.

$$\dot{a} = -k_a \frac{\partial \mathcal{F}}{\partial a} = -k_a \frac{\partial \mathcal{F}}{\partial s} \frac{\partial s}{\partial a} \tag{56}$$

This assumption is proposed to address a well-known issue in motor systems and control theory, namely the redundancy of effective movements (Franklin, 2011). Inverting a forward model to determine the action or policy (sequence of action) responsible for a given observation [**?**] can often lead to an ill-posed problem, as there can be multiple possible actions that could have generated a single observation. In active inference, this inverse model is divided into two sub-problems based on intrinsic (bodily) and extrinsic (environmental) frames of reference [**?**]. In the intrinsic frame, it is suggested that a significant portion of the control problem can be solved by predicting proprioceptive sensations in a similar manner to exteroceptive sensations. This involves using observations and a generative model of their dynamics to estimate the state of proprioceptive sensations (in Eq. 56 the term $\frac{\partial \mathcal{F}}{\partial s}$). On the other hand, the extrinsic problem in external coordinates is addressed by establishing simple heuristic mappings between proprioceptive estimates and observations, which can be implemented

as low-level reflexes (in Eq. 56 the term $\frac{\partial s}{\partial a}$). . This separation allows the generative model to handle the majority of the control task, including the generation of proprioceptive state predictions, while reflex arcs are assumed to be solved by agents over an evolutionary timescale [**?**]. This perspective heavily relies on proprioception, which is often studied and assumed to exist only in complex organisms. However, the concept of simple reflexes driving behavior can also be applicable to simpler organisms, where proprioceptive predictions could be attributed to even basic chemical networks triggering reflexes, such as tumbling in bacteria.

One could argue that Eq. 56 still represents an inverse model, as it aims to find the appropriate action for a desired output. However, unlike traditional approaches, active inference does not involve a mapping from hidden states x to actions a. Instead, it is formulated in terms of sensory data s directly. This implementation aligns with sensorimotor accounts of agent-environment systems, where action is fundamentally grounded in an extrinsic frame of reference [**?**], i.e., the real world s, rather than an intrinsic one based on inferred hidden states x obtained by inverting an internal forward model.