**Free Energy Workshop - 28th of October:**
From the free energy principle to experimental neuroscience, and back

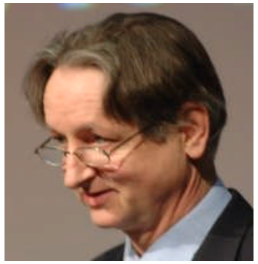
# The Bayesian brain, surprise and free-energy

Karl Friston, Wellcome Centre for Neuroimaging, UCL

## Abstract

Value-learning and perceptual learning have been an important focus over the past decade, attracting the concerted attention of experimental psychologists, neurobiologists and the machine learning community. Despite some formal connections; e.g., the role of prediction error in optimizing some function of sensory states, both fields have developed their own rhetoric and postulates. In work, we show that perception is, literally, an integral part of value learning; in the sense that it is necessary to integrate out dependencies on the inferred causes of sensory information. This enables the value of sensory trajectories to be optimized through action. Furthermore, we show that acting to optimize value and perception are two aspects of exactly the same principle; namely the minimization of a quantity (free-energy) that bounds the probability of sensations, given a particular agent or phenotype. This principle can be derived, in a straightforward way, from the very existence of biological agents, by considering the probabilistic behavior of an ensemble of agents belonging to the same class. Put simply, we sample the world to maximize the evidence for our existence

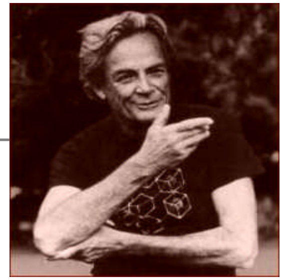

"Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism" - Hermann Ludwig Ferdinand von Helmholtz



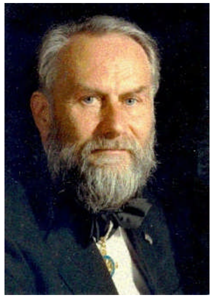
Geoffrey Hinton

Thomas Bayes

From the Helmholtz machine to the Bayesian brain and self-organization



Richard Feynman

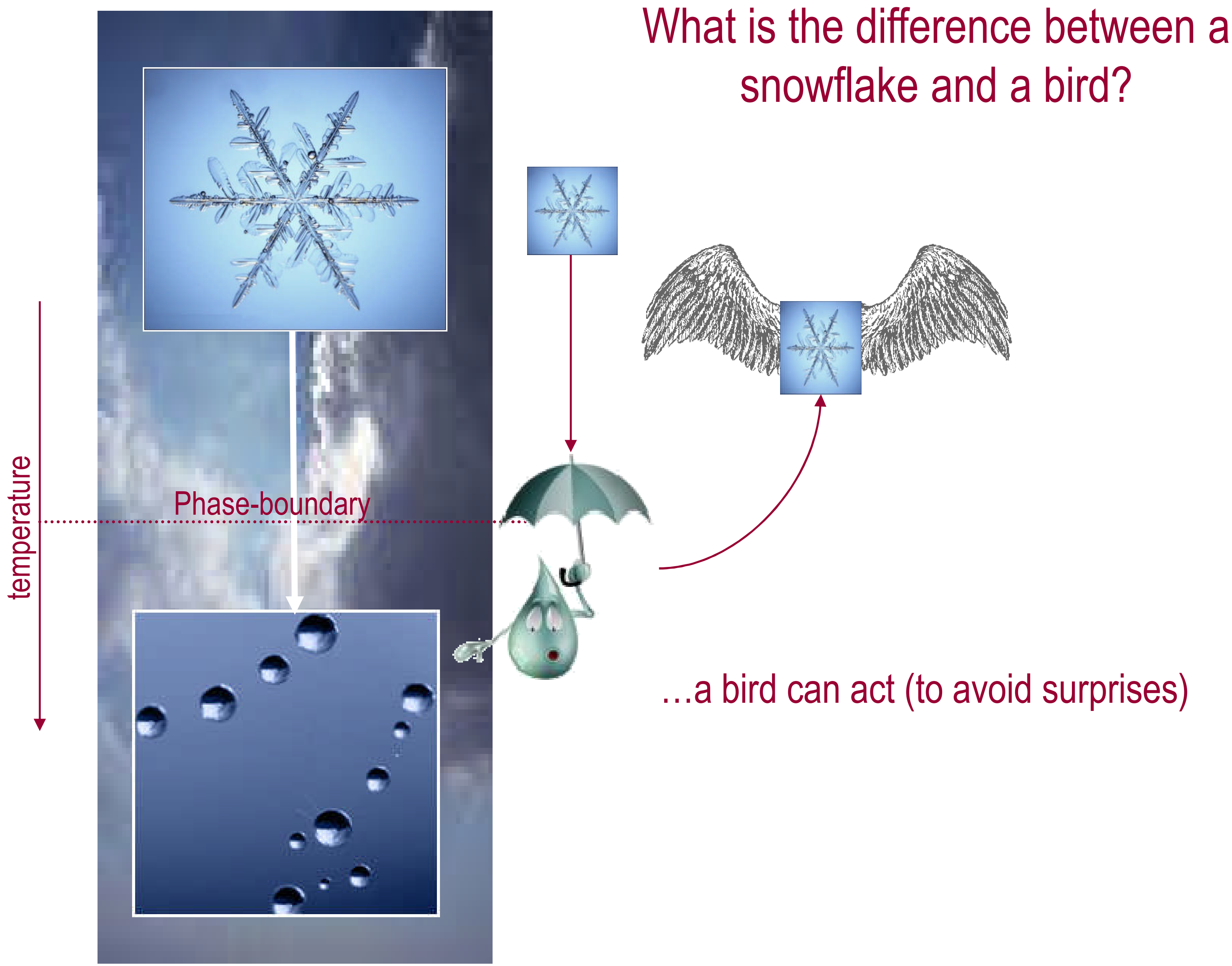


Hermann Haken

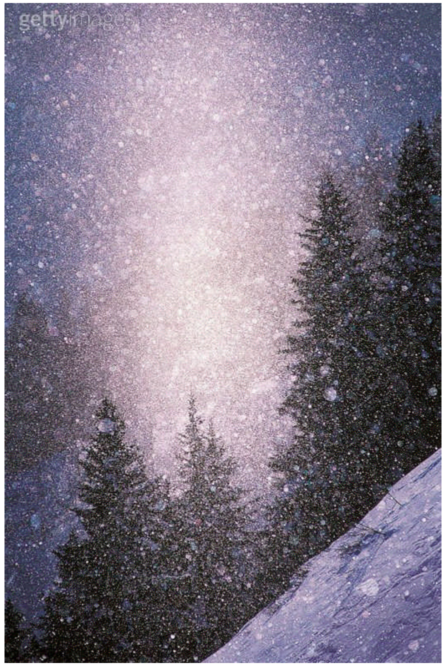
# Overview

| | |
|---|---|
| Ensemble dynamics | Entropy and equilibria<br>Free-energy and surprise |
| The free-energy principle | Action and perception<br>Hierarchies and generative models |

| | |
|---|---|
| Perception | Birdsong and categorization<br>Simulated lesions |
| Action | Active inference<br>Goal directed reaching |
| Policies | Control and attractors<br>The mountain-car problem |


What is the difference between a snowflake and a bird?
temperature
Phase-boundary
…a bird can act (to avoid surprises)

# What is the difference between snowfall and a flock of birds?



Ensemble dynamics, clumping and swarming

…birds (biological agents) stay in the same place

They resist the second law of thermodynamics, which says that their entropy should increase

# But what is the entropy?

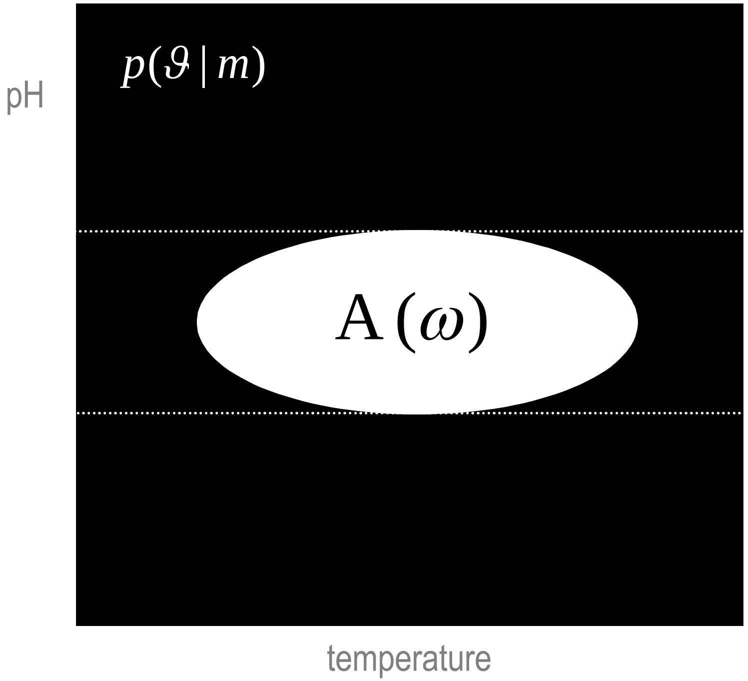
...entropy is just average surprise $\mathrm{H} = \int_0^T dt \mathrm{L}(t) \quad \mathrm{L}(t) = -\ln p(s \mid m)$

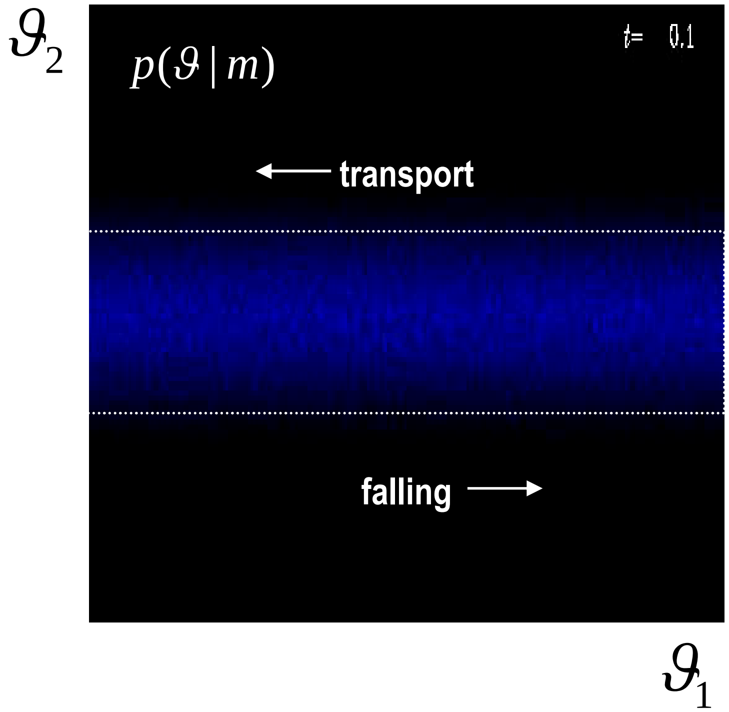
$s = \mathbf{g}(\vartheta)$


Low surprise (I am usually here)

High surprise (I am never here)

This means biological agents must self-organize to minimise surprise. In other words, to ensure they occupy a limited number of (attracting) states.




$$H = \int_0^T dt L(t) \propto -\int p(\vartheta \mid m) \ln p(\vartheta \mid m) d\vartheta$$

Self-organization that minimises the entropy of an ensemble density to ensure a limited repertoire of states are occupied (i.e., ensuring states have a random attracting set; cf homeostasis).

Particle density contours showing Kelvin-Helmholtz instability, forming beautiful breaking waves. In the self-sustained state of Kelvin-Helmholtz turbulence the particles are transported away from the mid-plane at the same rate as they fall, but the particle density is nevertheless very clumpy because of a clumping instability that is caused by the dependence of the particle velocity on the local solids-to-gas ratio (Johansen, Henning, & Klahr 2006)

But there is a small problem… agents cannot measure their surprise

$$s = \mathbf{g}(\vartheta)$$

?

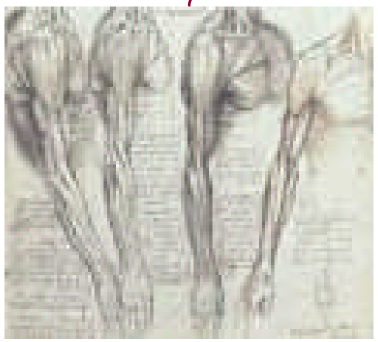
But they can measure their free-energy, which is always bigger than surprise

$$\mathrm{F}(t) \geq \mathrm{L}(t)$$

This means agents should minimize their free-energy. So what is free-energy?

# What is free-energy?

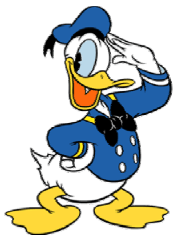
…free-energy is basically prediction error



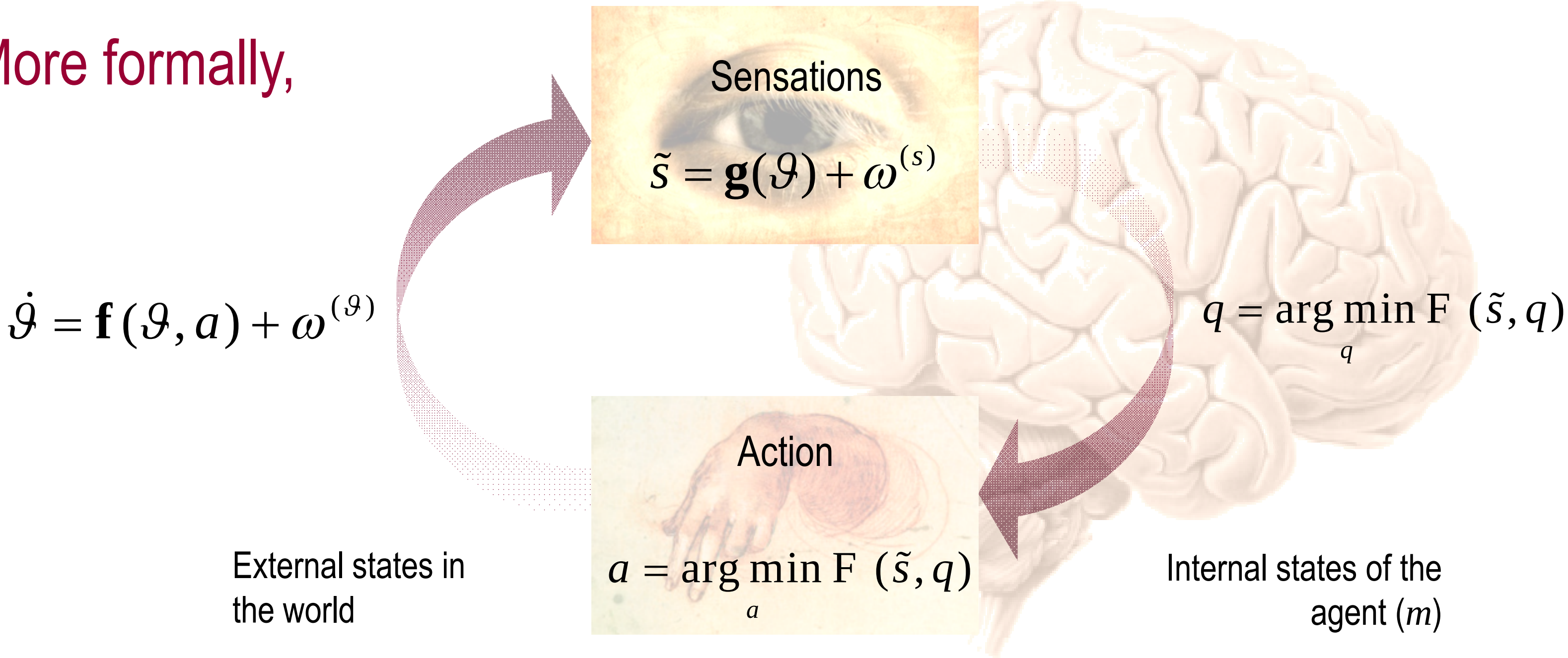
sensations – predictions

**= prediction error**

where small errors mean low surprise



# More formally,



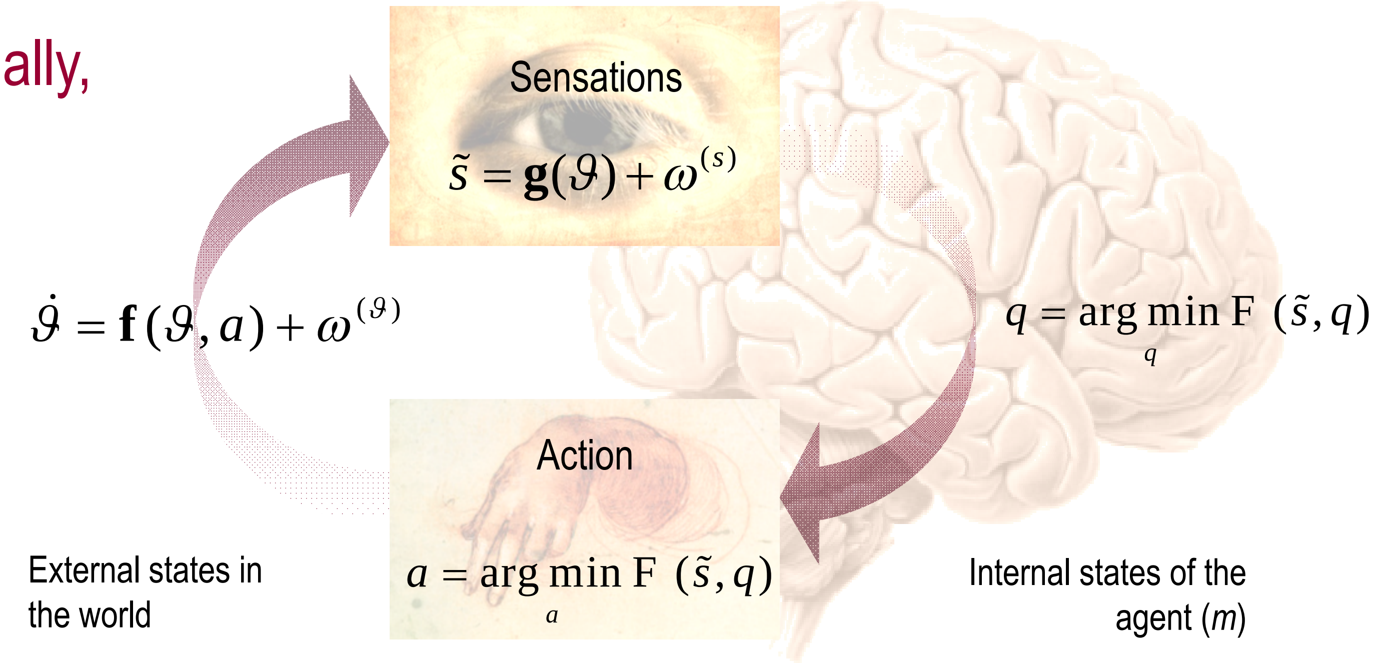

**Action to minimise a bound on surprise**

$$
\begin{aligned}
F &= D(q \,\|\, p(\vartheta)) - \langle \ln p(\tilde{s}(a) \,|\, \vartheta, m) \rangle_q \\
&= Complexity - Accuracy \\
a &= \arg\max_a Accuracy
\end{aligned}
$$

**Perception to optimise the bound**

$$
\begin{aligned}
F &= D(q(\vartheta \,|\, \mu) \,\|\, p(\vartheta \,|\, \tilde{s})) - \ln p(\tilde{s} \,|\, m) \\
&= Divergence + Surprise \\
\mu &= \arg\min_\mu Divergence
\end{aligned}
$$

# More formally,




Free-energy is a function of sensations and a recognition density over hidden causes

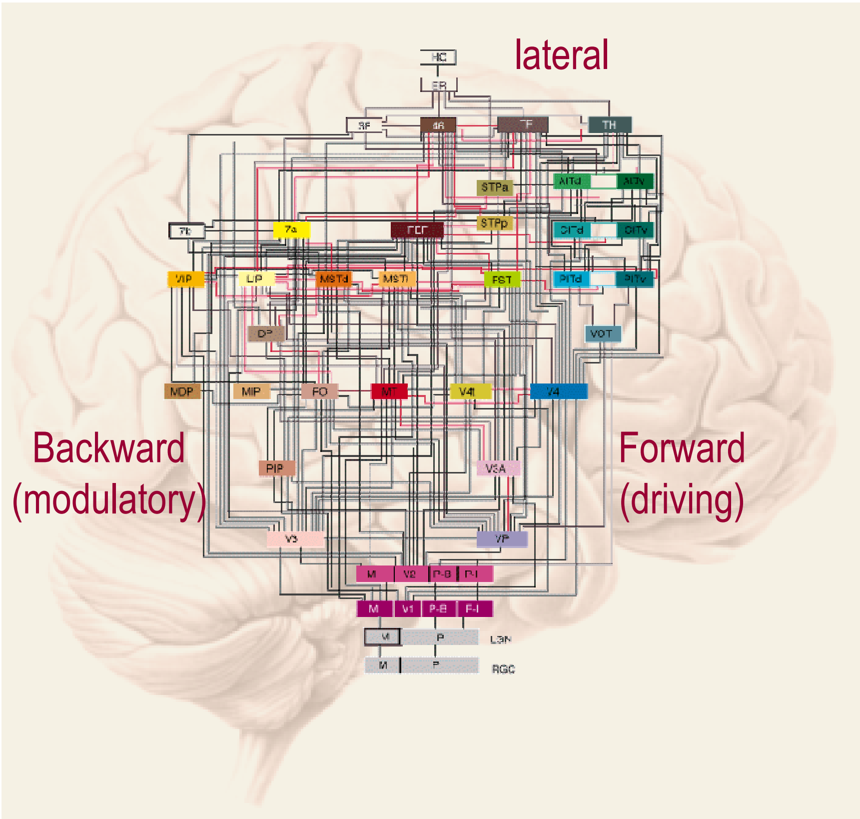
$$F(s, q) = Energy - Entropy = \langle G \rangle_q + \langle \ln q \rangle_q$$

and can be evaluated, given a generative model comprising a likelihood and prior:

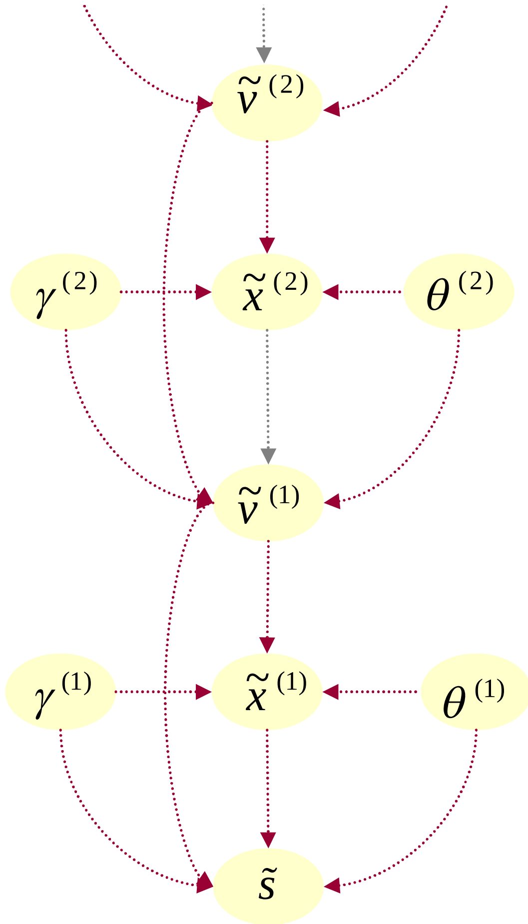
$$G(s, \vartheta) = -\ln p(s, \vartheta \mid m) = -\ln p(s \mid \vartheta, m) - \ln p(\vartheta \mid m)$$

So what models might the brain use?

# Hierarchal models in the brain




$$\tilde{v}^{(i-1)} = \tilde{g}^{(i)} + \tilde{\omega}^{(v,i)}$$
$$D\tilde{x}^{(i)} = \tilde{f}^{(i)} + \tilde{\omega}^{(x,i)}$$




$$\vartheta \supset \{x(t), v(t), \theta, \gamma\}$$

$$\vartheta \supset \{x(t), v(t), \theta, \gamma\}$$

## Hierarchal form

$$\begin{aligned} s &= g(x^{(1)}, v^{(1)}) + \omega^{(v,1)} \\ \dot{x}^{(1)} &= f(x^{(1)}, v^{(1)}) + \omega^{(x,1)} \\ &\vdots \\ v^{(i-1)} &= g(x^{(i)}, v^{(i)}) + \omega^{(v,i)} \\ \dot{x}^{(i)} &= f(x^{(i)}, v^{(i)}) + \omega^{(x,i)} \\ &\vdots \end{aligned}$$

## Gibb's energy: a simple function of prediction error

$$\begin{aligned} G &= -\ln p(\tilde{s}, \tilde{x}, \tilde{v} \mid m) \\ &= \tfrac{1}{2}\tilde{\varepsilon}^T \tilde{\Pi} \tilde{\varepsilon} - \tfrac{1}{2}\ln\left|\tilde{\Pi}\right| \end{aligned}$$

## Prediction errors

$$\varepsilon^{(v)} = \begin{bmatrix} s \\ v^{(1)} \\ \vdots \\ v^{(m)} \end{bmatrix} - \begin{bmatrix} g^{(1)} \\ \vdots \\ g^{(m)} \\ \eta \end{bmatrix}$$

$$\tilde{\varepsilon} = \begin{bmatrix} \tilde{\varepsilon}^{(v)} = \tilde{v} - \tilde{g} \\ \tilde{\varepsilon}^{(x)} = D\tilde{x} - \tilde{f} \end{bmatrix}$$

## Likelihood and empirical priors

$$p(\tilde{s}, \tilde{x}, \tilde{v} \mid m) = p(\tilde{s} \mid \tilde{x}, \tilde{v})\, p(\tilde{x}, \tilde{v})$$

$$p(\tilde{x}, \tilde{v}) = \prod_{i=1}^{n-1} p(x^{(i)})\, p(D\tilde{x}^{(i)} \mid \tilde{v}^{(i)})\, p(\tilde{v}^{(i)} \mid \tilde{x}^{(i+1)}, \tilde{v}^{(i+1)})$$

$p\left(D\tilde{x}^{(i)} \mid x^{(i)}, \tilde{v}^{(i)}\right) = N(\tilde{f}^{(i)}, \Pi^{(x,i)})$ Dynamical priors

$p\left(\tilde{v}^{(i)} \mid \tilde{x}^{(i+1)}, \tilde{v}^{(i+1)}\right) = N(\tilde{g}^{(i+1)}, \Pi^{(v,i)})$ Structural priors

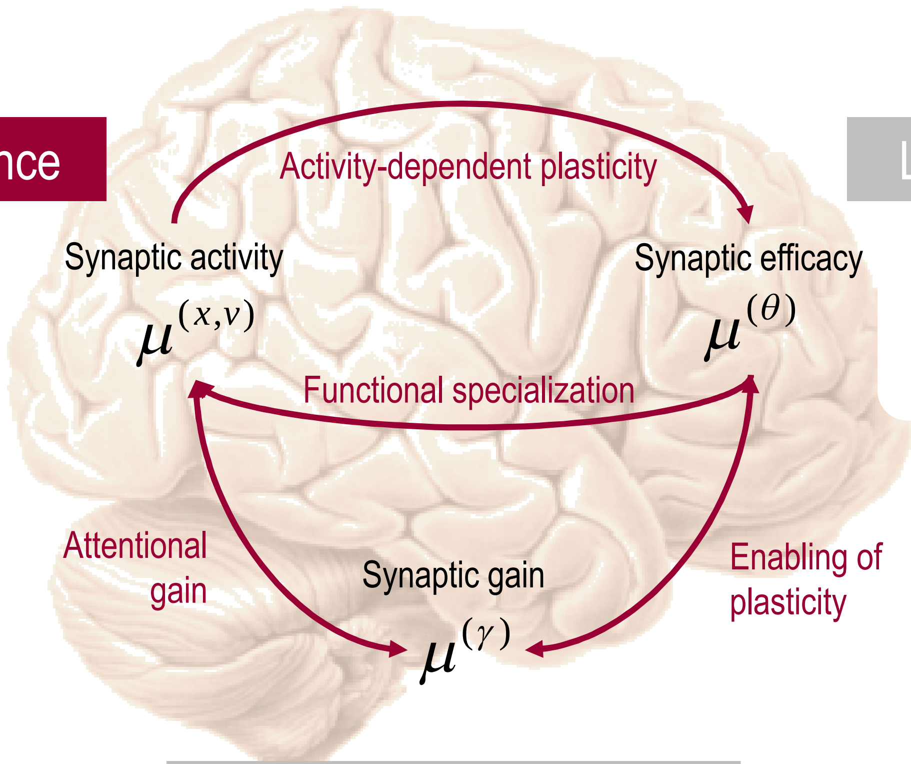
# The recognition density and its sufficient statistics

Laplace approximation: $q(\vartheta \mid \mu) = N\,(\mu, \Sigma(\mu))$

## Perception and inference

$$\dot{\mu}^{(x)} = -\frac{\partial G}{\partial \mu^{(x)}}$$

$$\dot{\mu}^{(v)} = -\frac{\partial G}{\partial \mu^{(v)}}$$



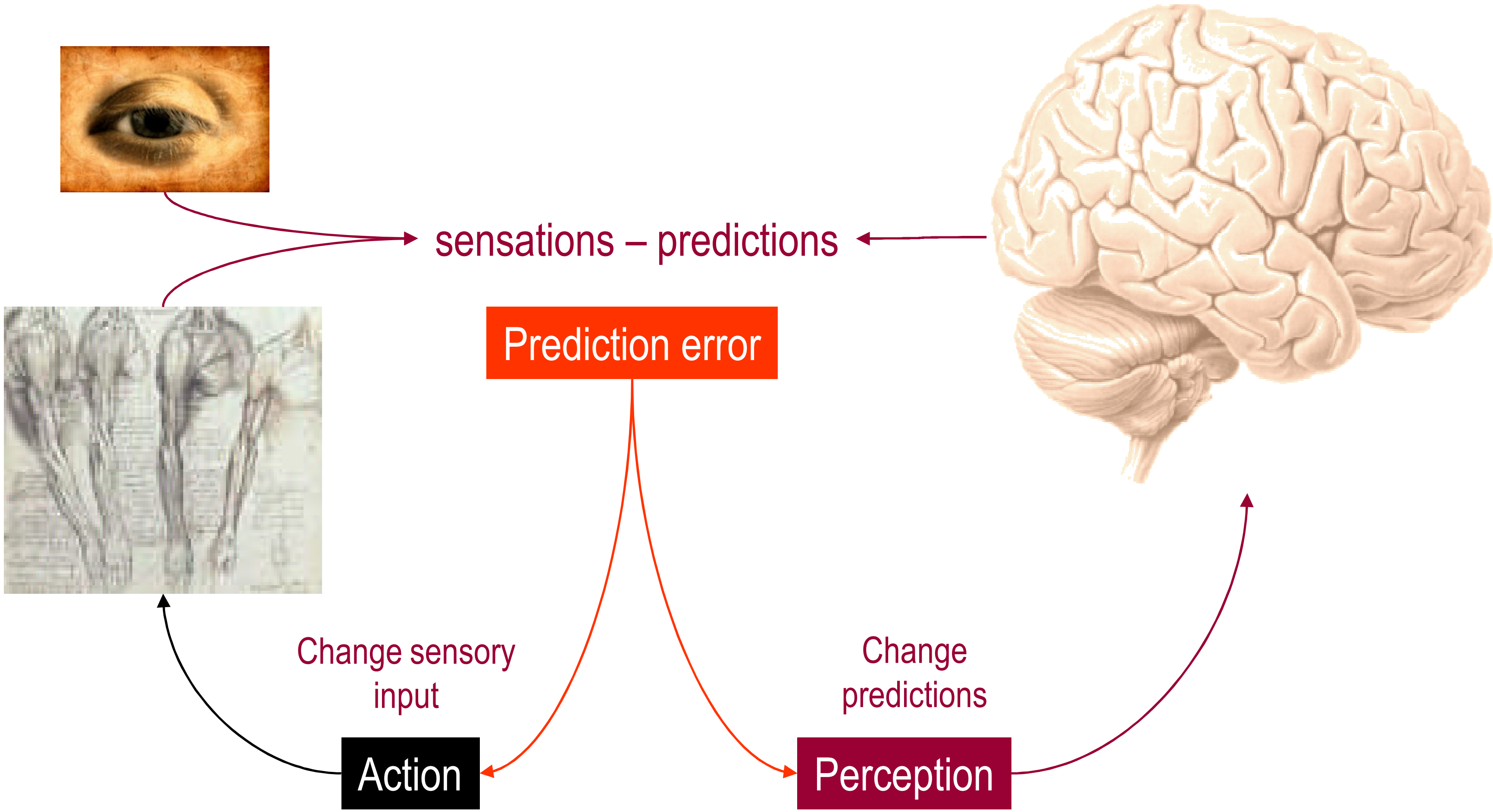

## Learning and memory

$$\ddot{\mu}^{(\theta)} = -\frac{\partial G}{\partial \mu^{(\theta)}} - \kappa \dot{\mu}^{(\theta)}$$

## Attention and salience

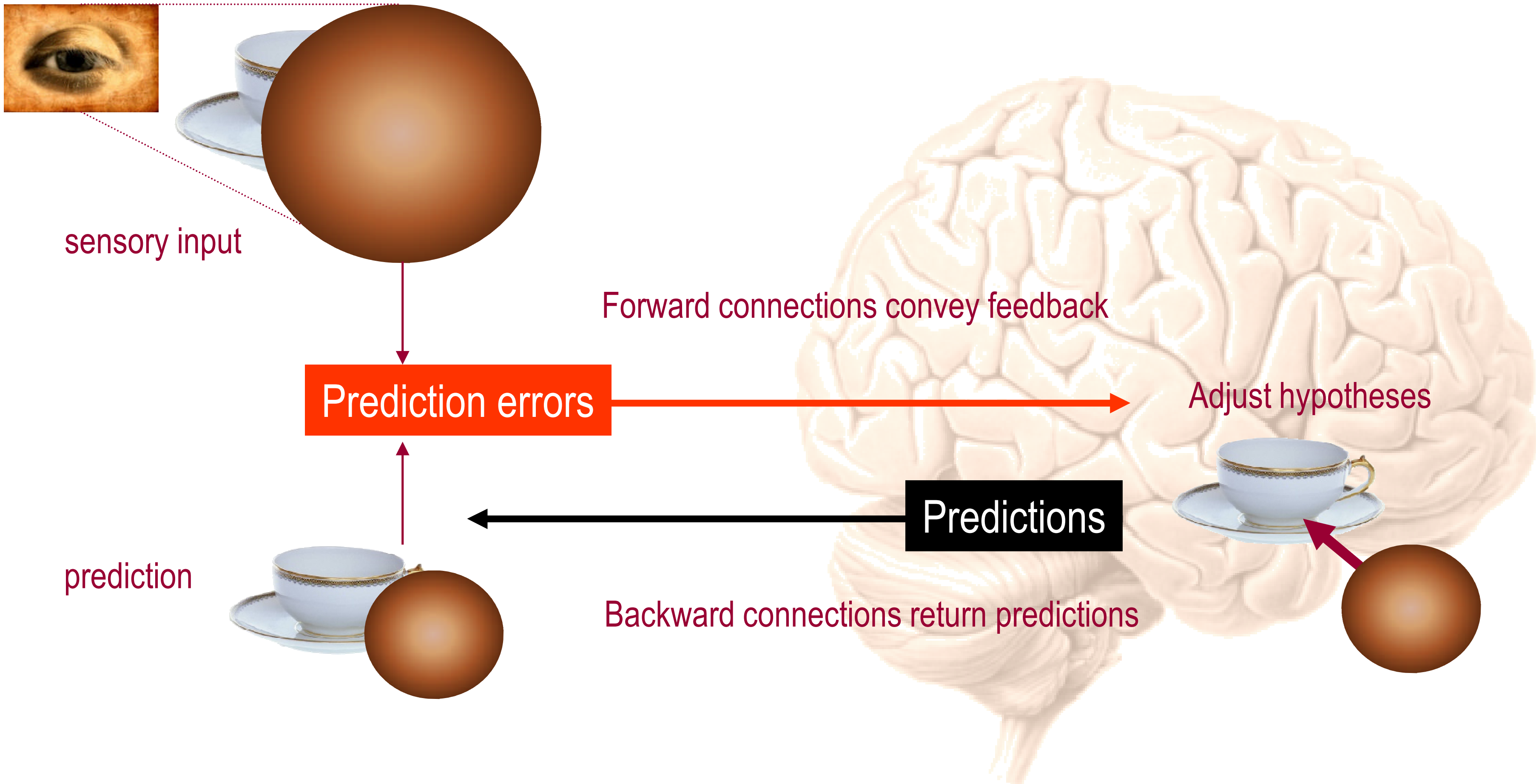
$$\ddot{\mu}^{(\gamma)} = -\frac{\partial G}{\partial \mu^{(\gamma)}} - \kappa \dot{\mu}^{(\gamma)}$$

# How can we minimize prediction error (free-energy)?




…prediction errors drive action and perception to suppress themselves

# So how do prediction errors change predictions?




...by hierarchical message passing in the brain

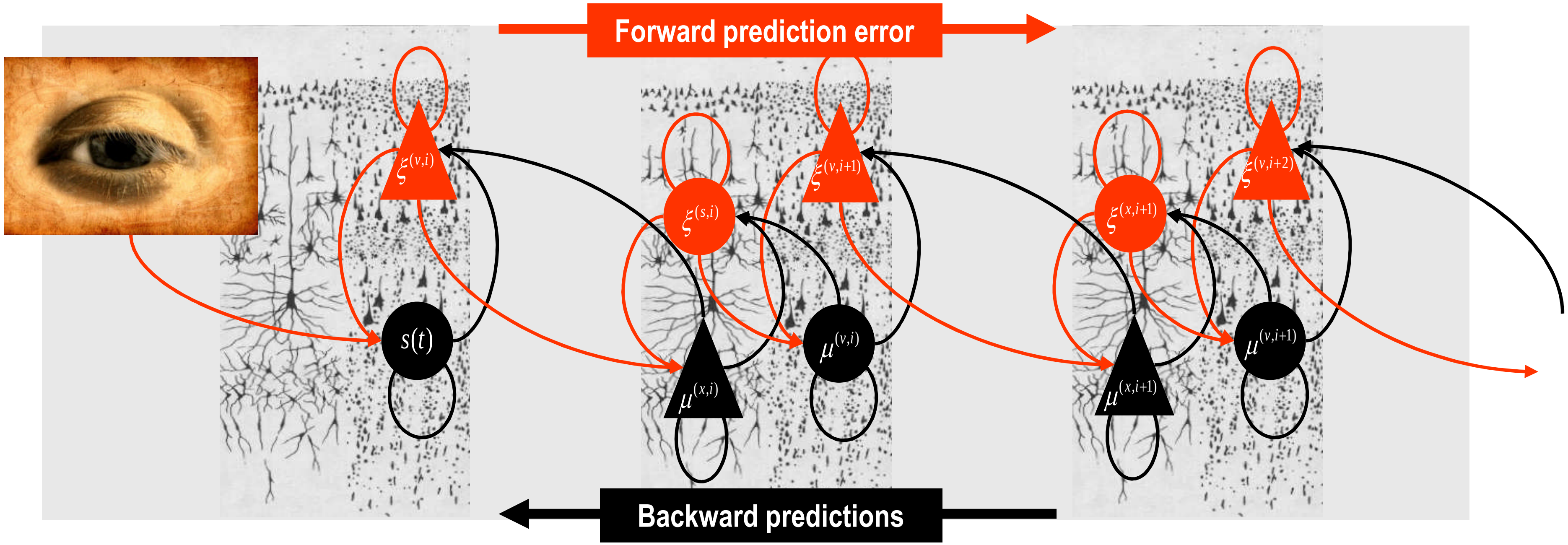
# More formally,

David Mumford

## Perception and message-passing

$$\xi^{(v,i)} = \Pi^{(v,i)} \tilde{\varepsilon}^{(v,i)} = \Pi^{(v,i)} (\tilde{\mu}^{(v,i-1)} - \tilde{g})$$

$$\xi^{(x,i)} = \Pi^{(x,i)} \tilde{\varepsilon}^{(x,i)} = \Pi^{(x,i)} (\mathrm{D}\tilde{\mu}^{(x,i)} - \tilde{f}^{(i)})$$




$$\dot{\tilde{\mu}}^{(v,i)} = \mathrm{D}\tilde{\mu}^{(v,i)} - \tilde{\varepsilon}_v^{(i)T} \xi^{(i)} - \xi^{(v,i+1)}$$

$$\dot{\tilde{\mu}}^{(x,i)} = \mathrm{D}\tilde{\mu}^{(x,i)} - \tilde{\varepsilon}_x^{(i)T} \xi^{(i)}$$

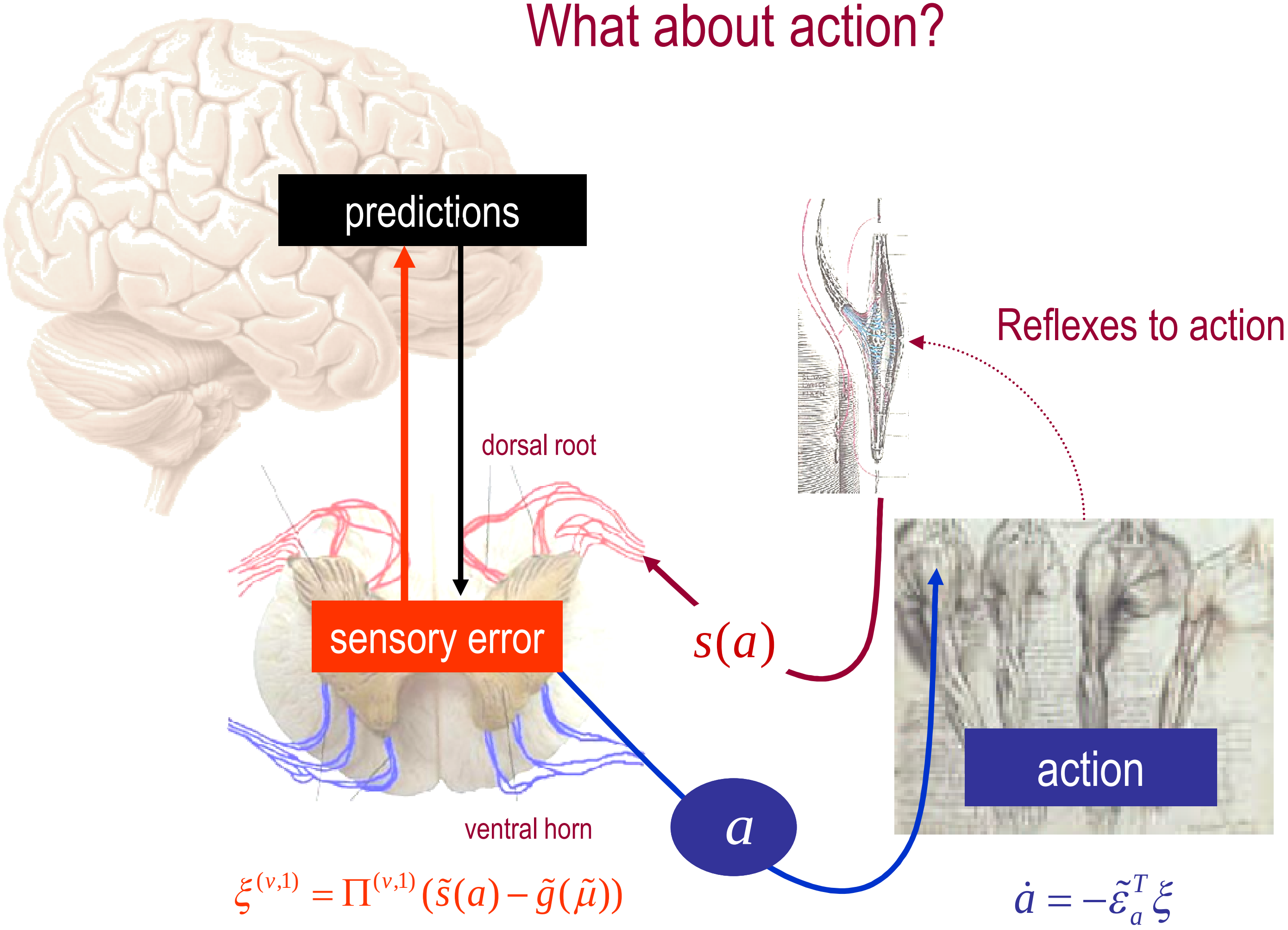
## Synaptic plasticity

$$\ddot{\mu}_i^{(\theta)} = -\tilde{\varepsilon}_{\theta_i}^T \xi$$

## Synaptic gain

$$\ddot{\mu}_i^{(\gamma)} = \tfrac{1}{2} tr(R_i(\xi\xi^T - \Pi(\mu^{(\gamma)})))$$

# What about action?




Action can only suppress (sensory) prediction error.
This means action fulfils our (sensory) predictions

# Summary

Biological agents resist the second law of thermodynamics

They must minimize their average surprise (entropy)

They minimize surprise by suppressing prediction error (free-energy)

Prediction error can be reduced by changing predictions (perception)

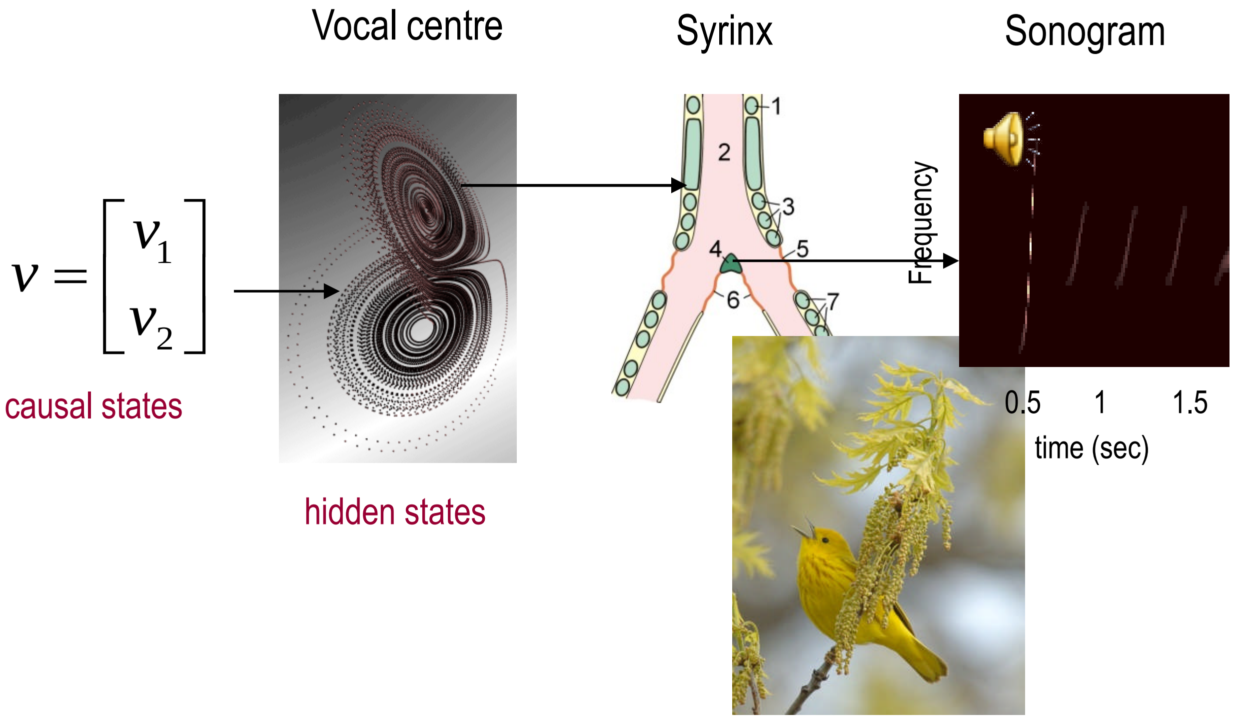
Prediction error can be reduced by changing sensations (action)

Perception entails recurrent message passing in the brain to optimise predictions

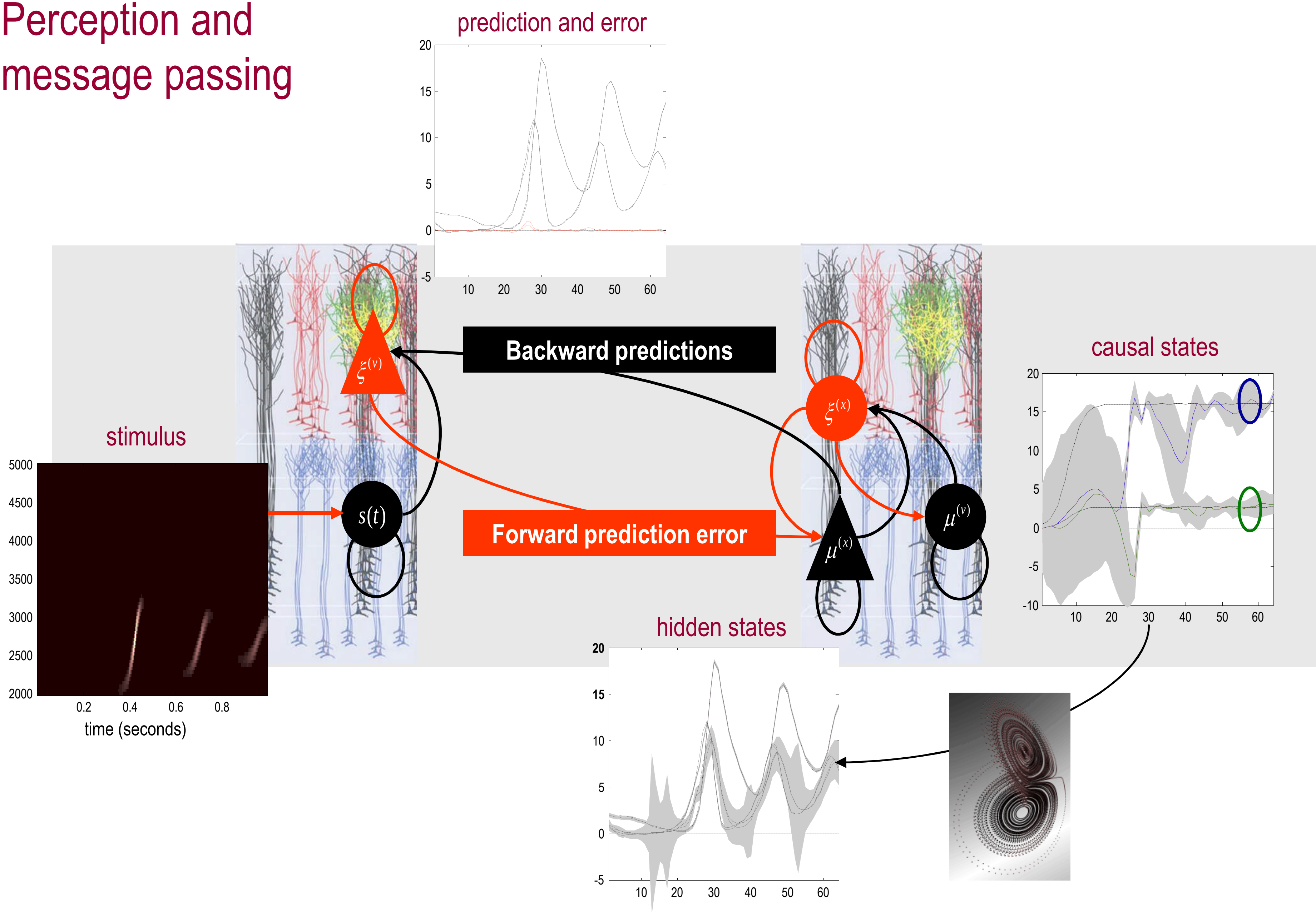
Action makes predictions come true (and minimises surprise)

| | |
|---|---|
| **Perception** | Birdsong and categorization<br>Simulated lesions |
| **Action** | Active inference<br>Goal directed reaching |
| **Policies** | Control and attractors<br>The mountain-car problem |

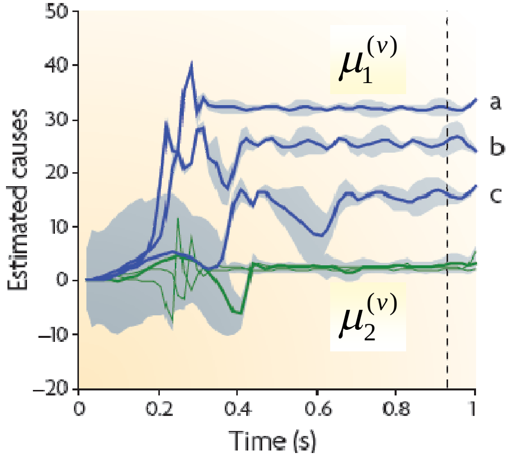
# Making bird songs with Lorenz attractors

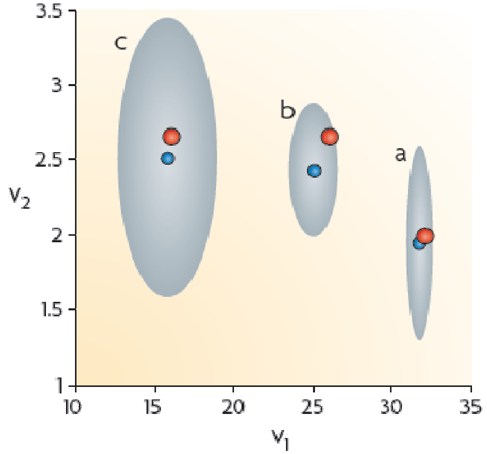


# Perception and message passing

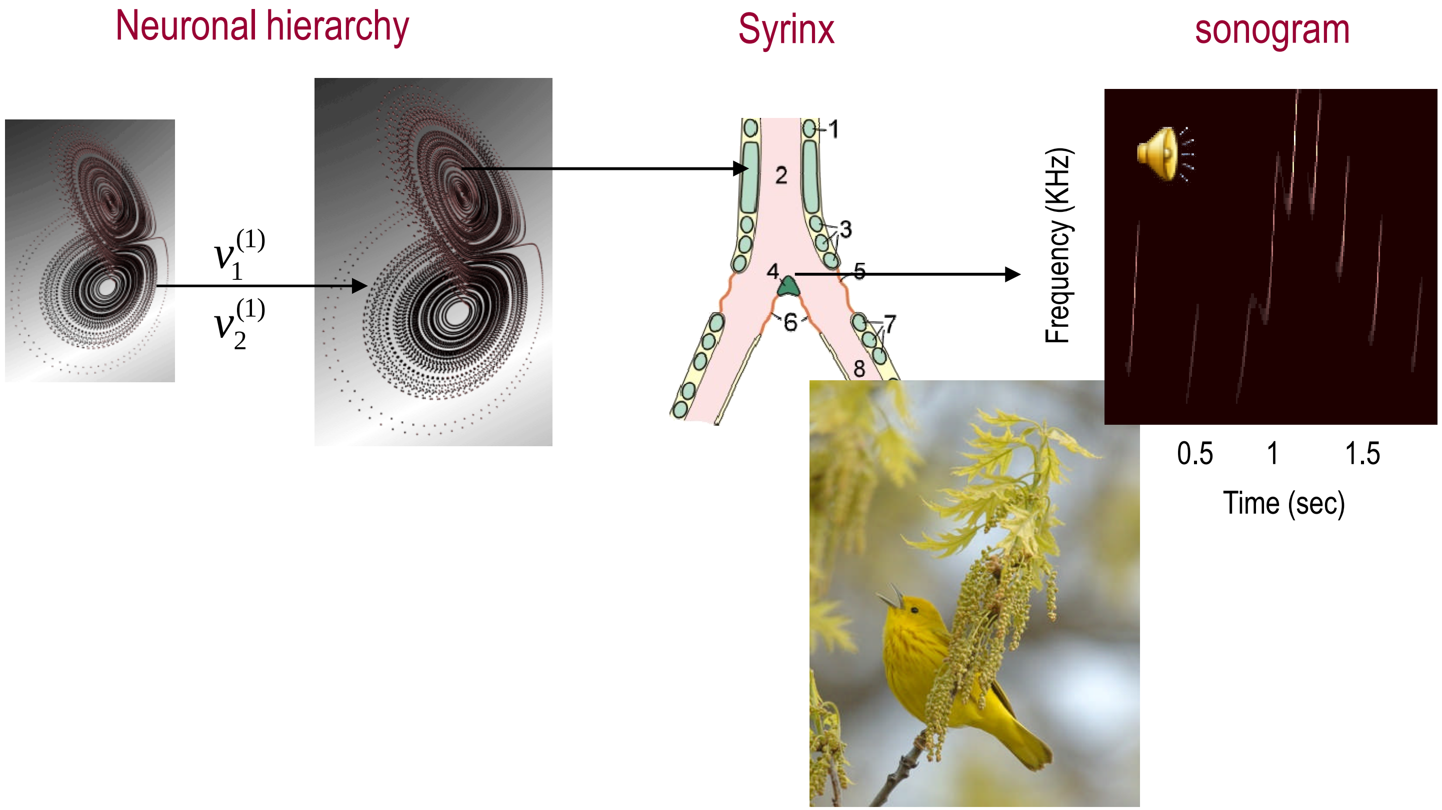


# Perceptual categorization

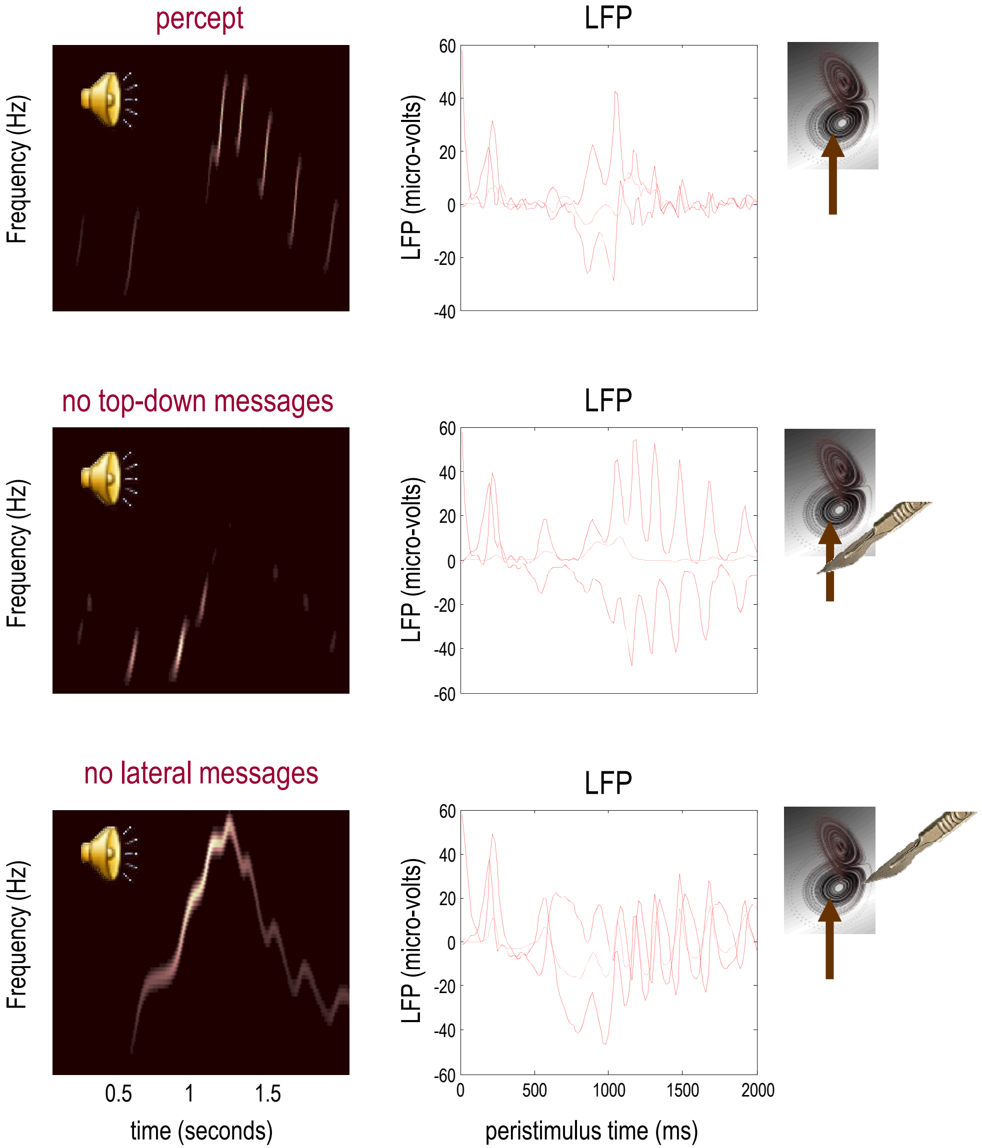
# Hierarchical (itinerant) birdsong: sequences of sequences



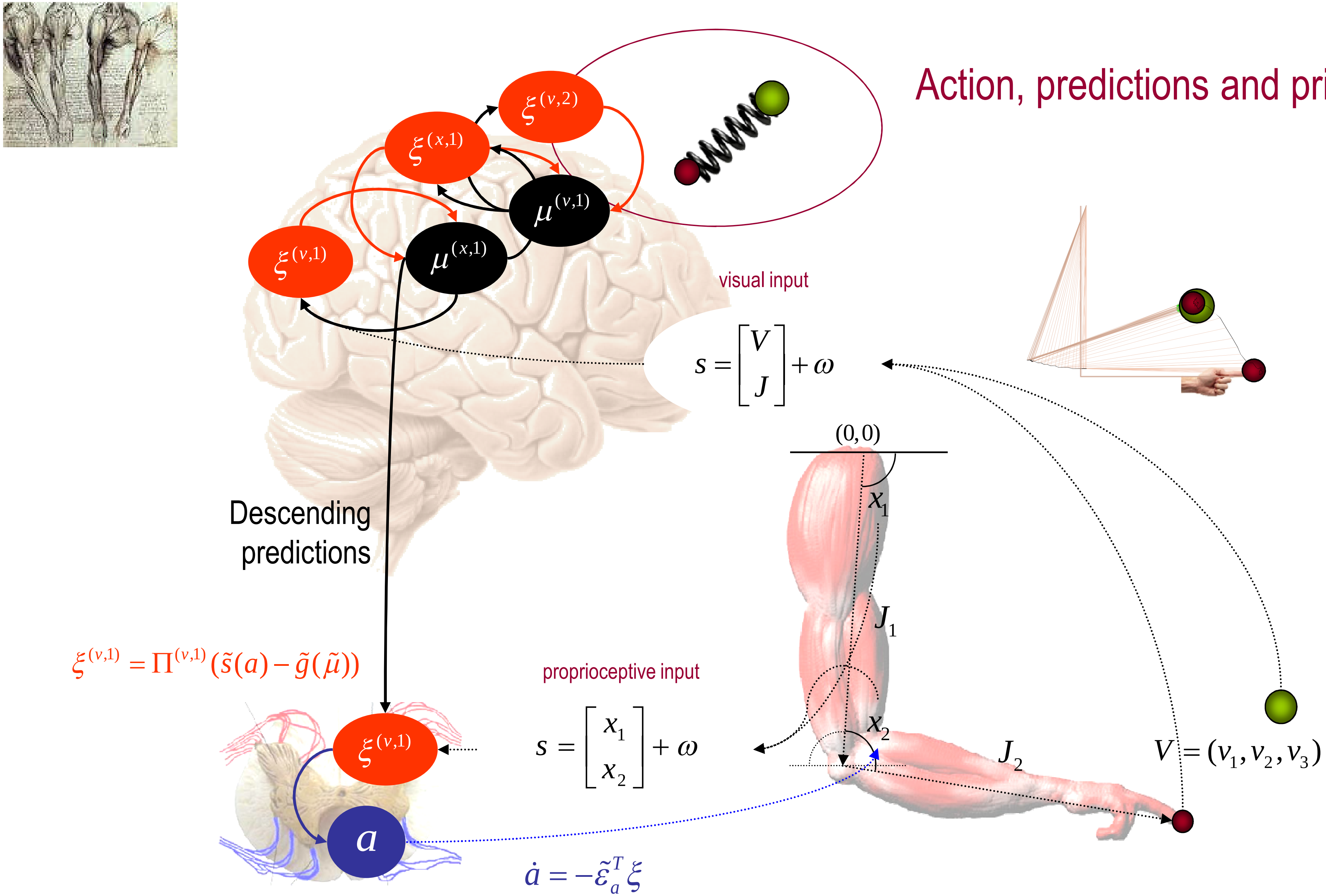
# Simulated lesions and hallucinations

no structural priors

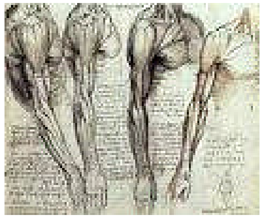
no dynamical priors

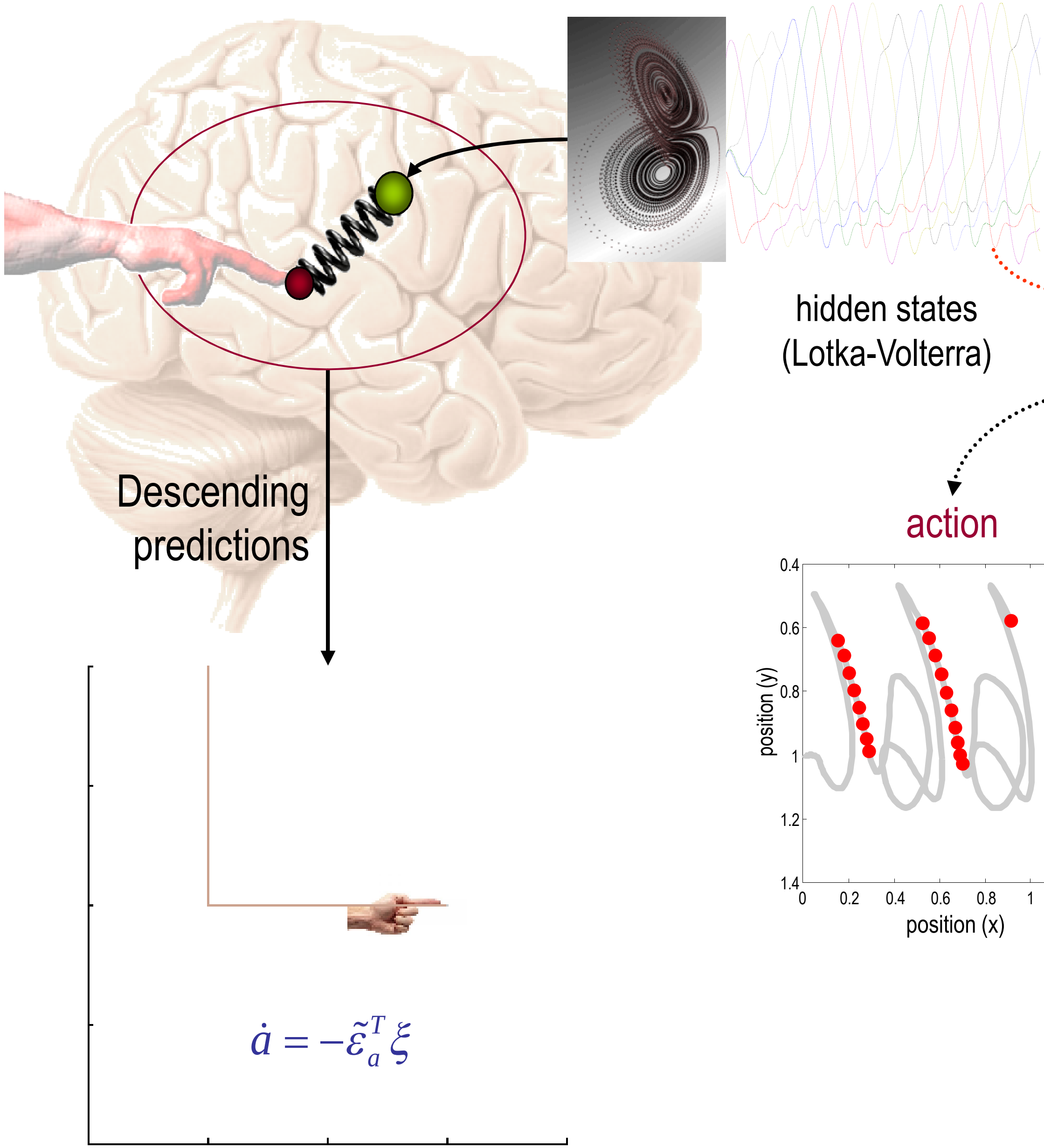


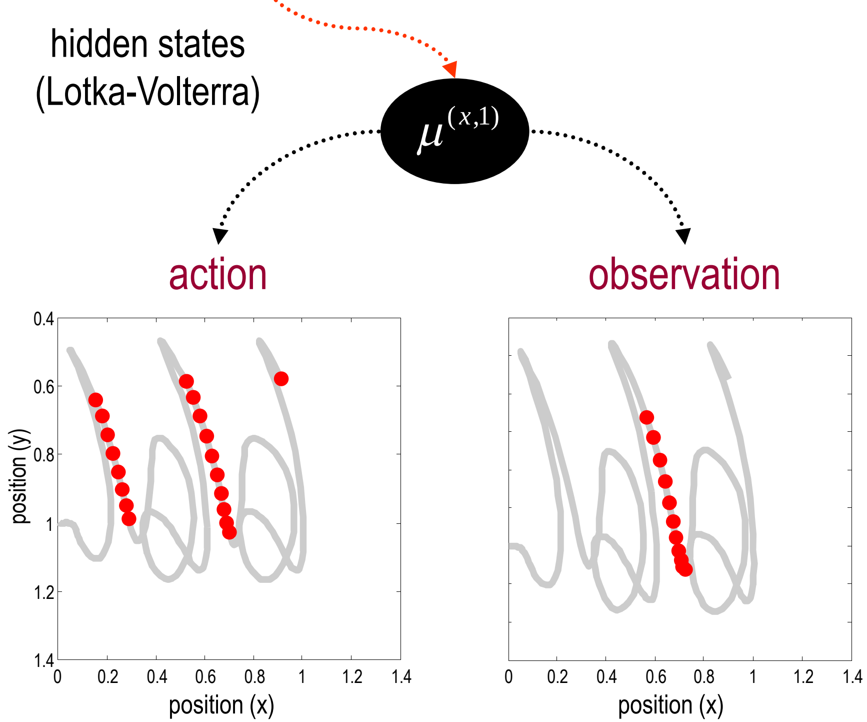
# Action, predictions and priors



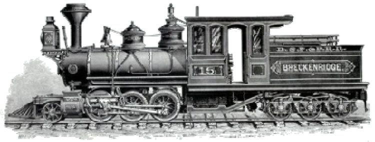
# Itinerant behavior and action-observation





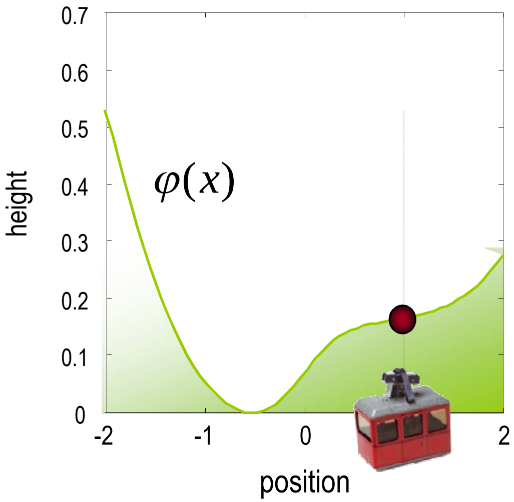
Adriaan Fokker



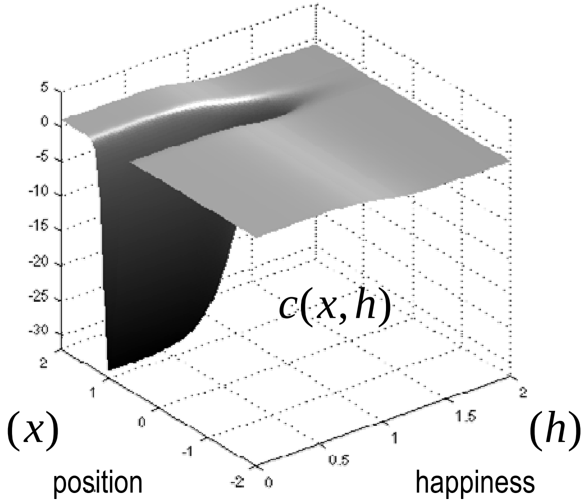
# The mountain car problem

Max Planck

## The environment




## The cost-function



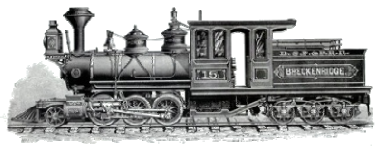

## True motion

$$\tilde{\mathbf{f}} = \begin{bmatrix} \dot{x} \\ \dot{x}' \end{bmatrix} = \begin{bmatrix} x' \\ \sigma(a) - \varphi_x - \frac{1}{8} x' \end{bmatrix}$$

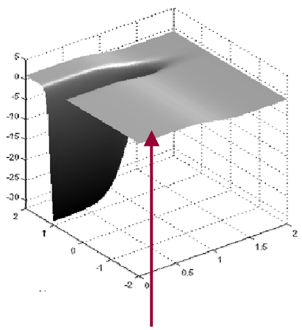
## Policy (predicted motion)

$$\tilde{f} = \begin{bmatrix} \dot{x} \\ \dot{x}' \end{bmatrix} = \begin{bmatrix} x' \\ cx' - \varphi_x \end{bmatrix}$$

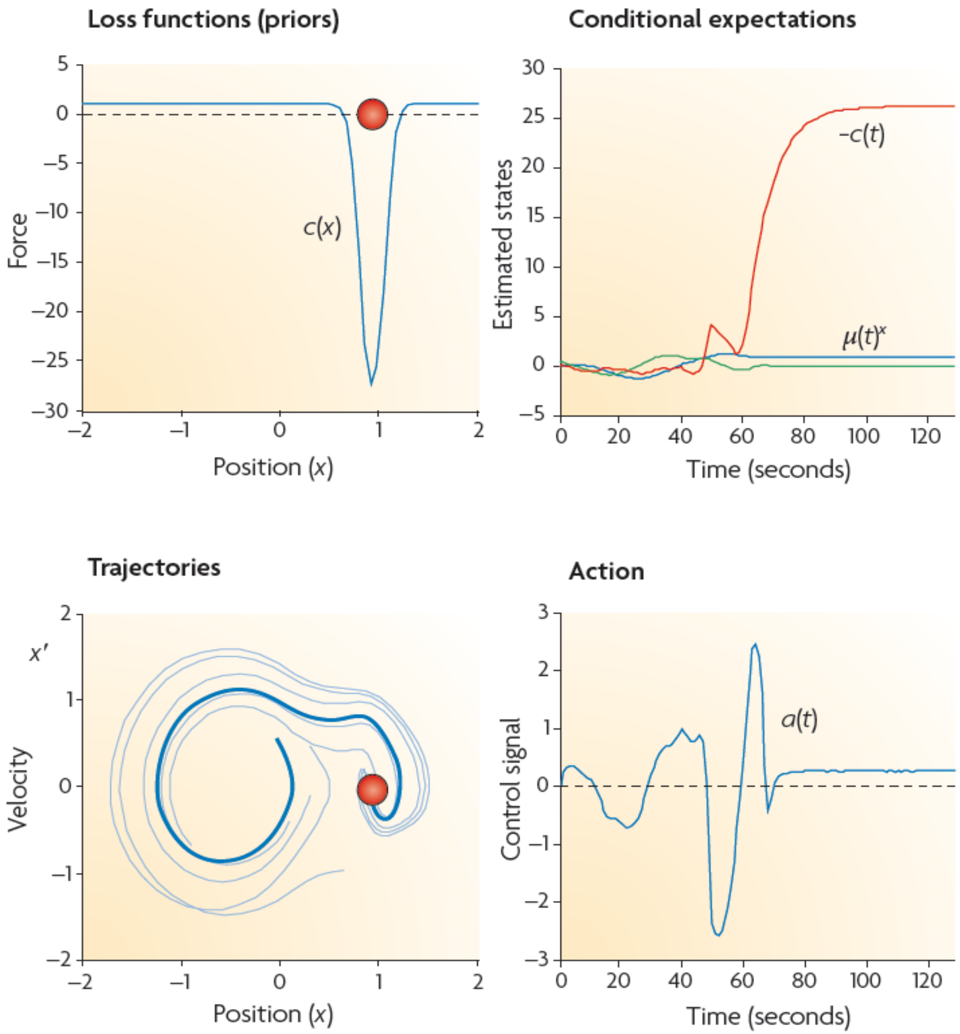
"I expect to move faster when cost is positive"

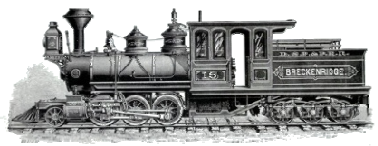


# Exploring & exploiting the environment

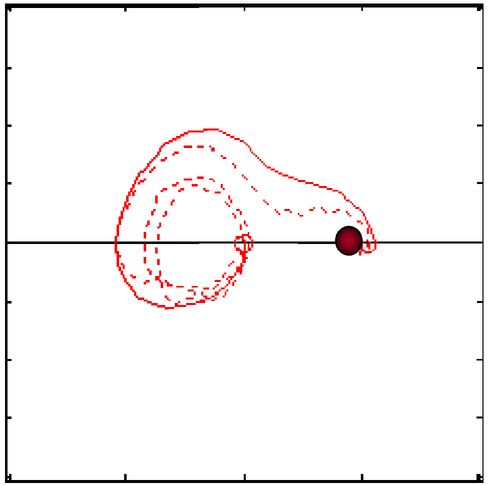


With cost
(i.e., exploratory
dynamics)





# Policies and prior expectations



Using just the free-energy principle and itinerant priors on motion, we have solved a benchmark problem in optimal control theory (without learning).

If priors are so important, where do they come from?

# …we inherit them

Darwinian evolution of virtual block creatures. A population of several hundred creatures is created within a supercomputer, and each creature is tested for their ability to perform a given task, such the ability to swim in a simulated water environment. The successful survive, and their virtual genes are copied, combined, and mutated to make offspring. The new creatures are again tested, and some may be improvements on their parents. As this cycle of variation and selection continues, creatures with more and more successful behaviours can emerge.

# The selection of adaptive predictions

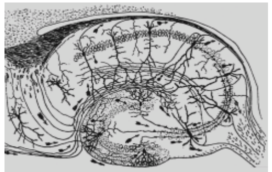
## Time-scale

## Free-energy minimisation leading to…

$10^{-3}\,s$

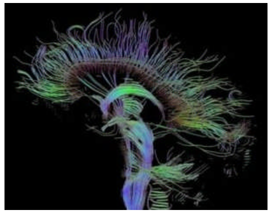
**Perception and Action:** The optimisation of neuronal and neuromuscular activity to suppress prediction errors (or free-energy) based on generative models of sensory data.

$10^{0}\,s$



**Learning and attention:** The optimisation of synaptic gain and efficacy over seconds to hours, to encode the precisions of prediction errors and causal structure in the sensorium. This entails suppression of free-energy over time.

$10^{3}\,s$



$10^{6}\,s$

**Neurodevelopment:** Model optimisation through activity-dependent pruning and maintenance of neuronal connections that are specified epigenetically

**Evolution:** Optimisation of the average free-energy (free-fitness) over time and individuals of a given class (e.g., conspecifics) by selective pressure on the epigenetic specification of their generative models.

$10^{15}\,s$

# Thank you

And thanks to collaborators:

Jean Daunizeau
Harriet Feldman
Lee Harrison
Stefan Kiebel
James Kilner
Jérémie Mattout
Klaas Stephan

And colleagues:

Peter Dayan
Jörn Diedrichsen
Paul Verschure
Florentin Wörgötter