heidelberg.ai

Oct. 9th, 7:30pm: How Neuroscience can help to solve AI (1/2) powered by

MIC

MEDICAL IMAGE COMPUTING

From the Bayesian Brain to Active Inference...

...and the other way round.

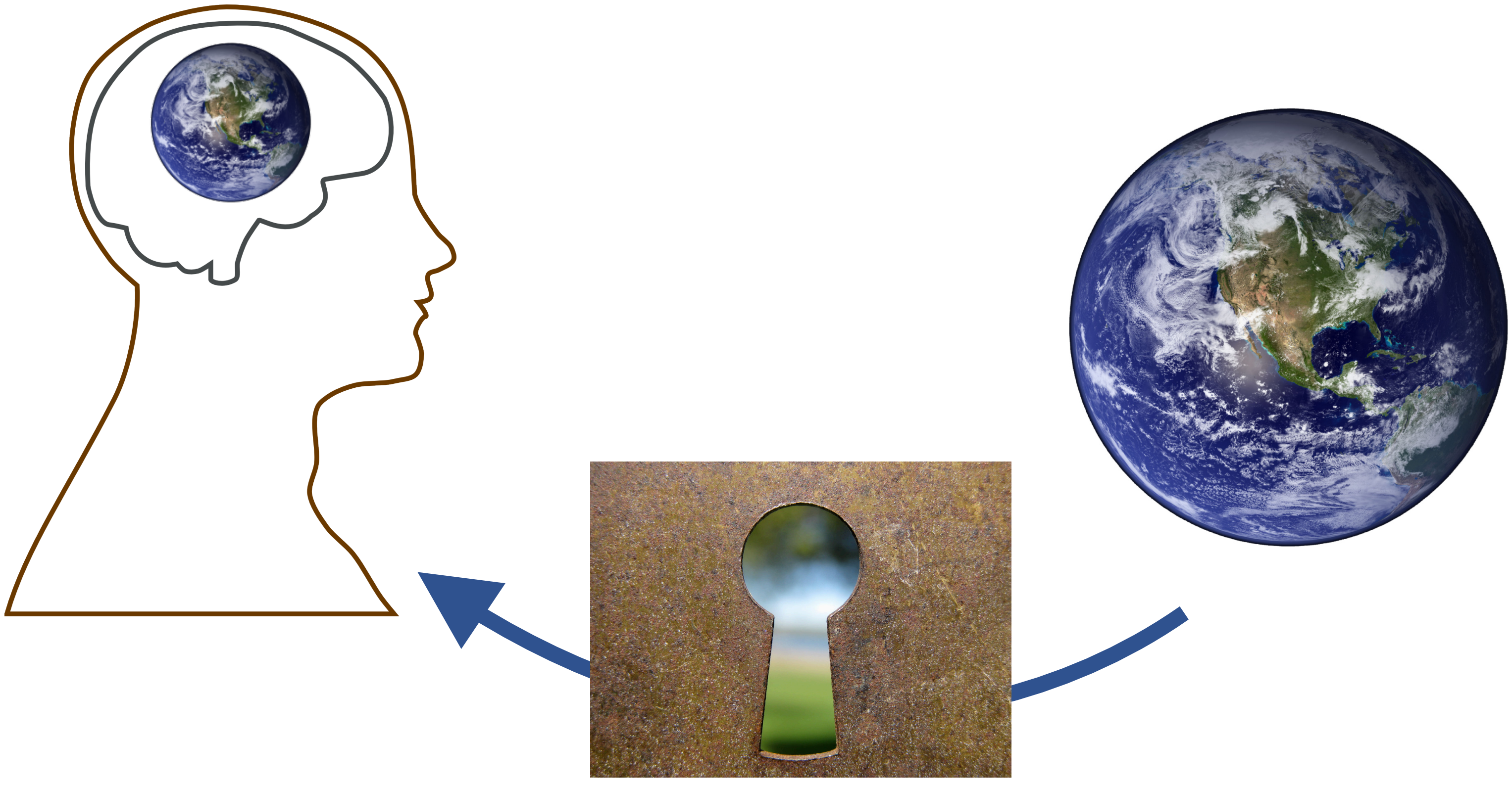
Kai Ueltzhöffer, 9.10.2017

Disclaimer

- Today:
  Overview Talk! 100% *NOT* my own work.
  But important to give some context and motivation for…

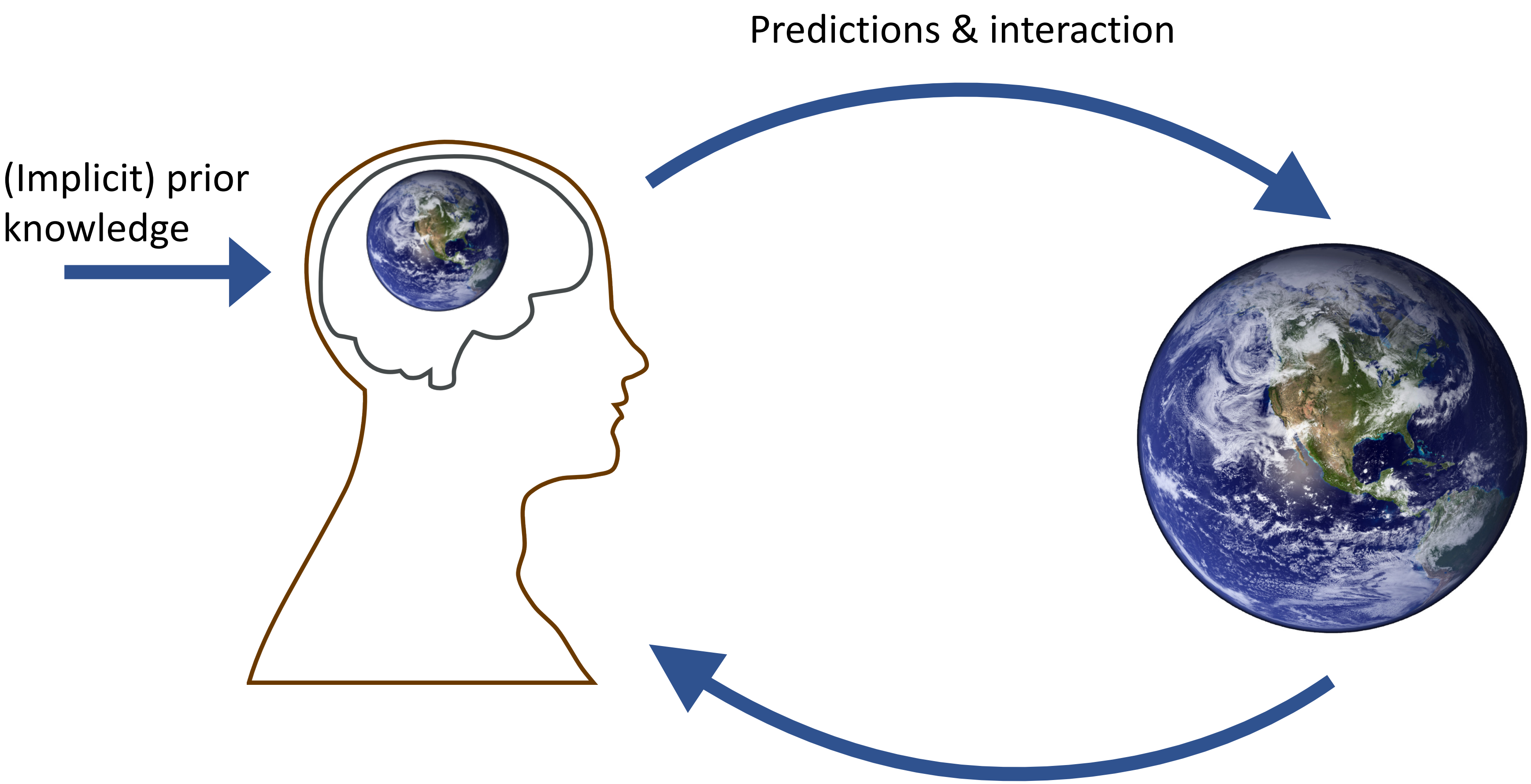
- Next week:
  Mostly my own work (+ some basics) ☺.

How do we perceive the world?



Senses: Vision, Hearing, Smell, Taste, Touch, Nociception, Interoception, Proprioception

A (possible) solution



Hermann von Helmholz,
"Handbuch der
physiologischen Optik",
1867

Senses: Vision, Hearing, Smell, Taste,
Touch, Interoception, Proprioception

How to formalise such a theory?

- Probability theory allows to make **exact statements** about **uncertain information**.
- Among others, a recipe to **optimally** combine **a priori knowledge** (“a prior”) with **observations**.
  → Bayes’ Theorem

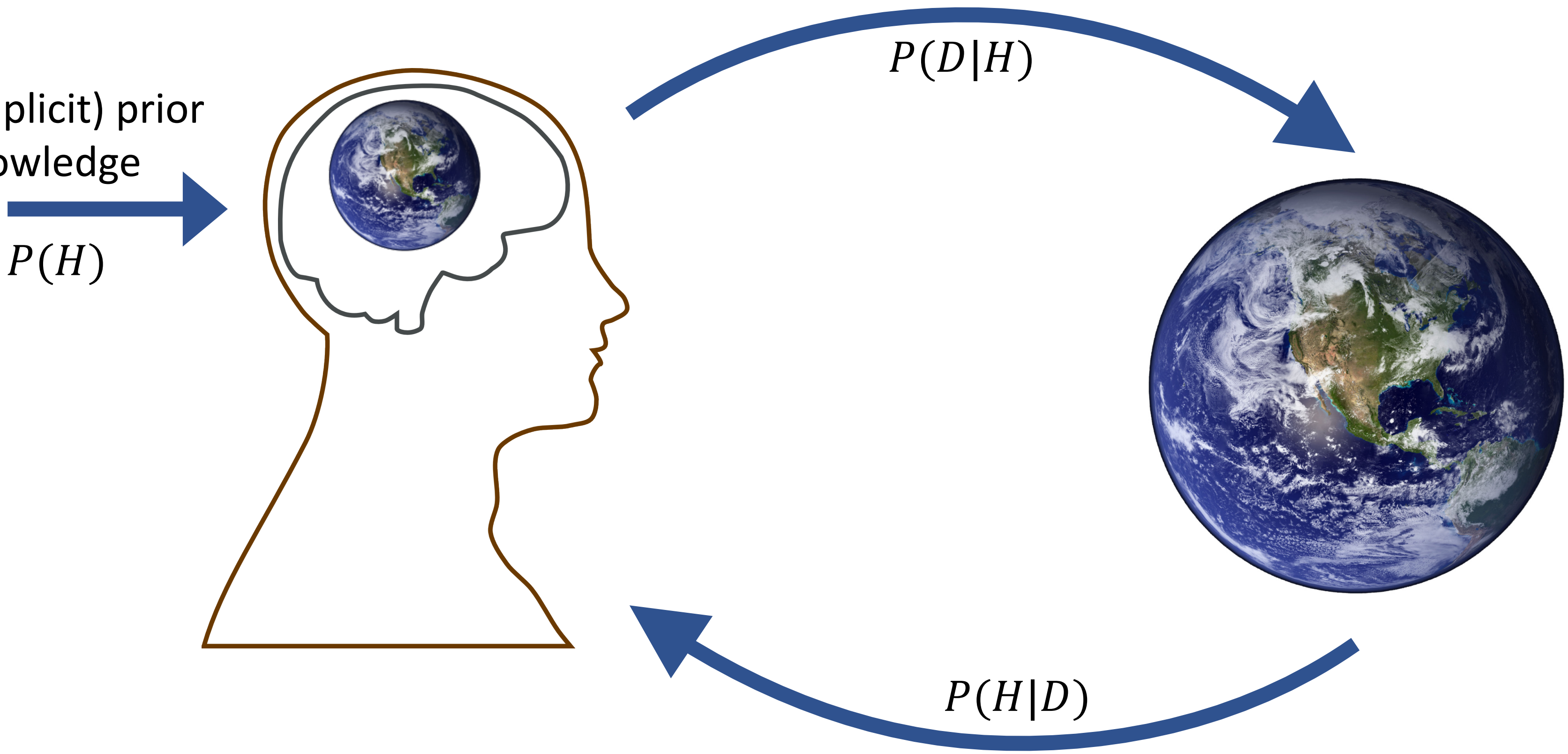
Bayes' Theorem

$$P(H|D)P(D) = P(H,D) = P(D|H)P(H)$$
$$\Longrightarrow P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

- P(H): "Prior" probability that hypothesis H about the world is true.
- P(D): Probability of observing D
- P(D|H): Probability of observing D, given that hypothesis H is true. → "Likelihood" function.
- P(H|D): Probability that hypothesis H is true, given that D was observed. → "Posterior"

Thomas Bayes,
1701-1761

A (possible) solution



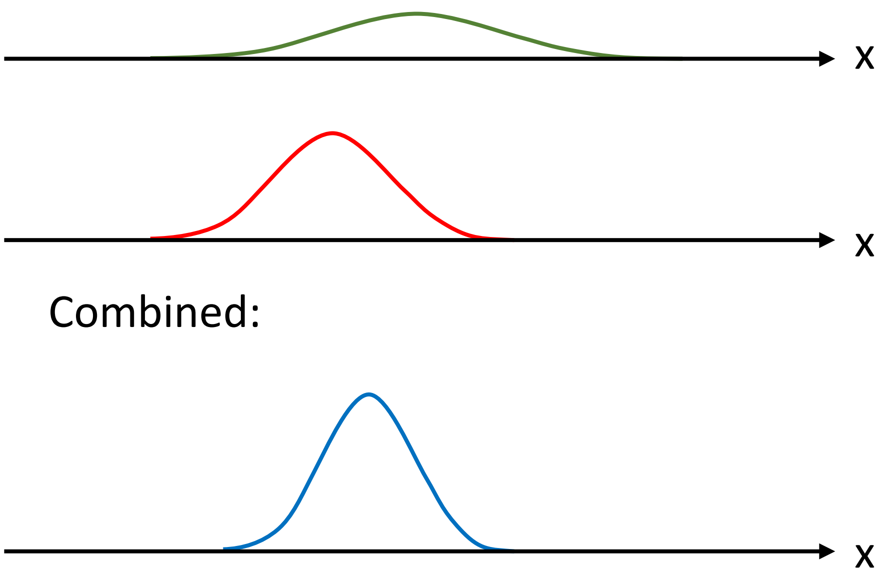
Hermann von Helmholtz, "Handbuch der physiologischen Optik", 1867

Senses: Vision, Hearing, Smell, Taste, Touch, Interoception, Proprioception

Optimal perception with Bayes' Theorem

"Tock, tock, tock, …"

$$P(X|A) = \frac{P(A|X)P(X)}{P(A)}$$



P(X): Prior probability for Hypothesis "The woodpecker* sits at position X". A woodpecker should be somewhere close to the trunk of the tree.

P(A|X): Probability of hearing "toc, toc, toc" from the left side of the tree, given the bird's position is X. Likelihood function allows to imagine sensory consequences from hypotheses about the world.

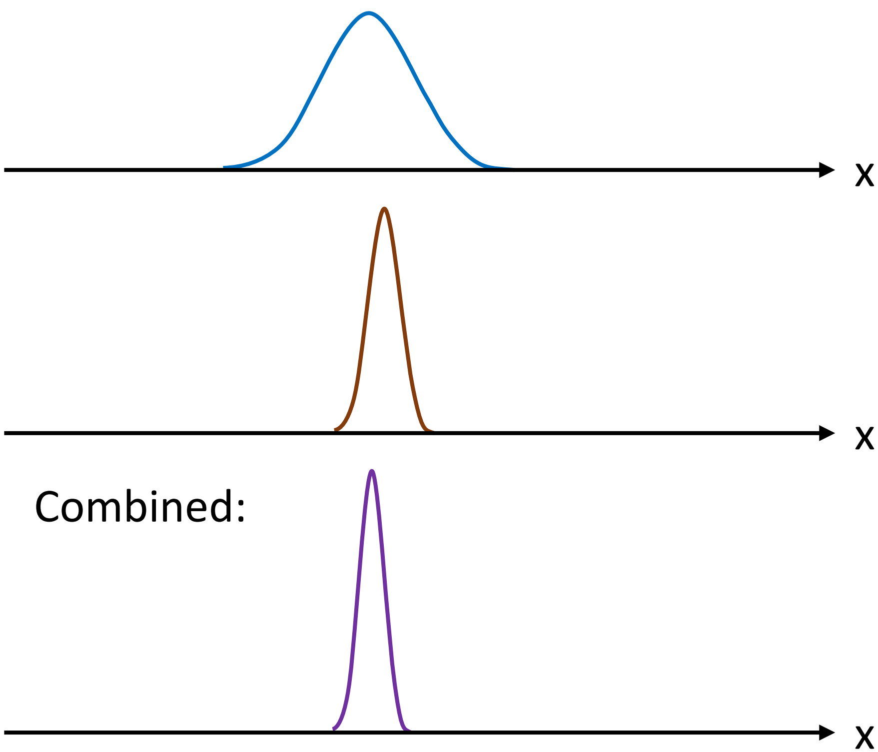
P(X|A): Posterior probability of the bird's position X, given the "toc, toc, toc" sound is heard at the let side of the tree.

*woodpecker = Specht

Optimal perception with Bayes' Theorem

"Tock, tock, tock, …"

$$P(H|A,V\ ) = \frac{P(V|X)P(X|A)}{P(V|A)}$$

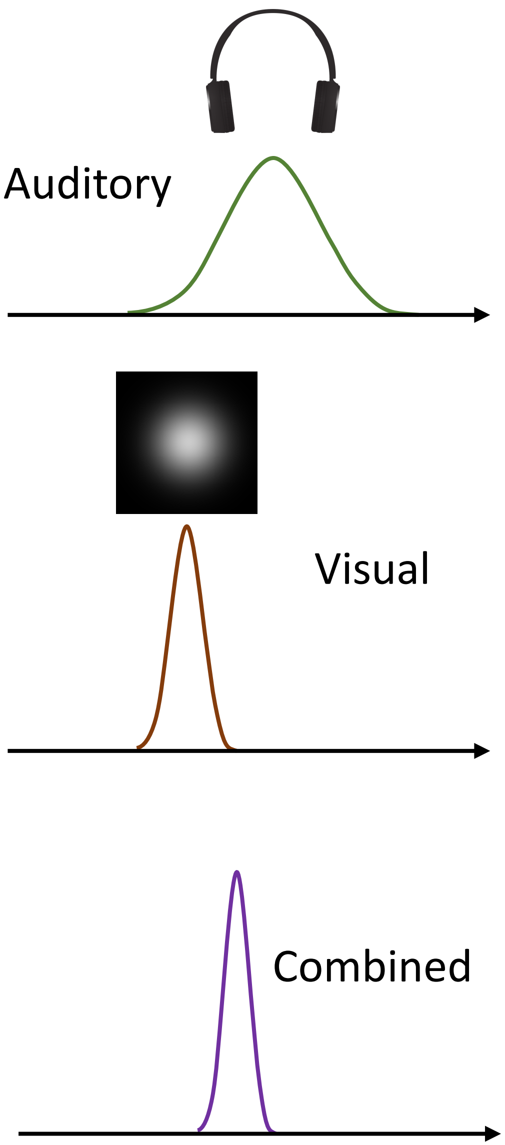


P(X|A): Posterior probability of the bird's position X, given the "toc, toc, toc" sound is heard at the let side of the tree.

P(V|X): Probability of observing the woodpecker at the left side of the trunk, given it's position X.

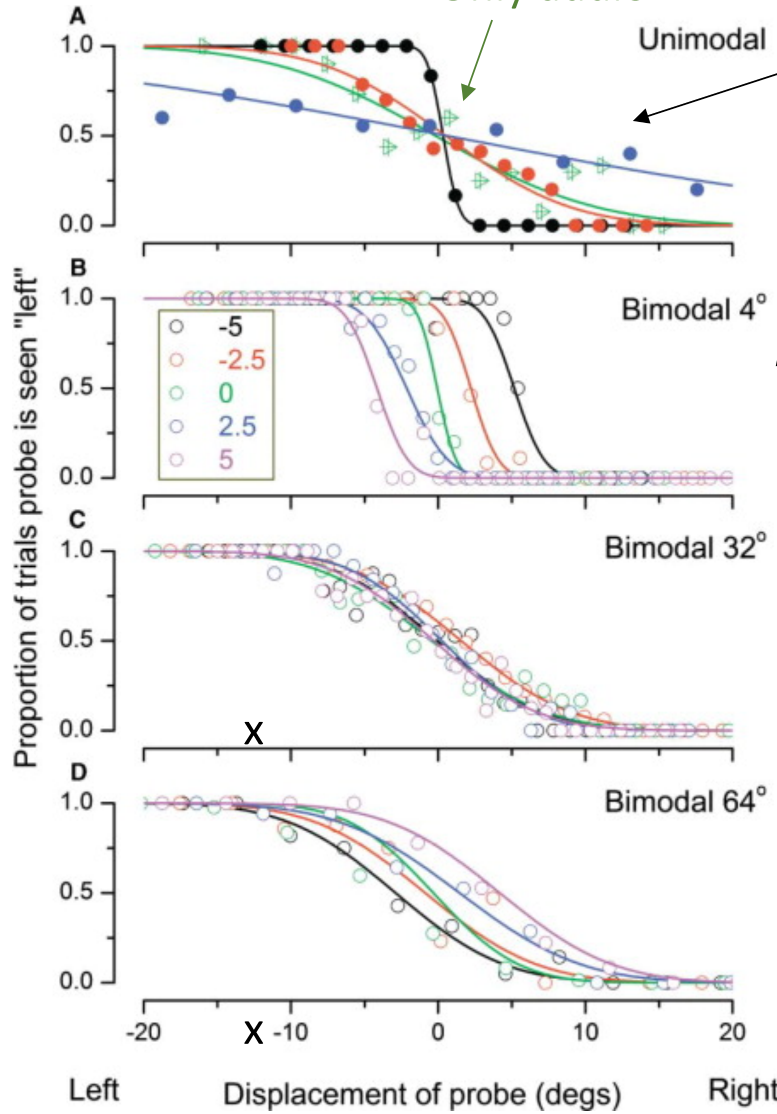
P(X|A,V): Posterior probability of the bird's position X, given auditory and visual information.

Sounds reasonable, but might it be true?



Only audio

Only visual information with decreasing accuracy

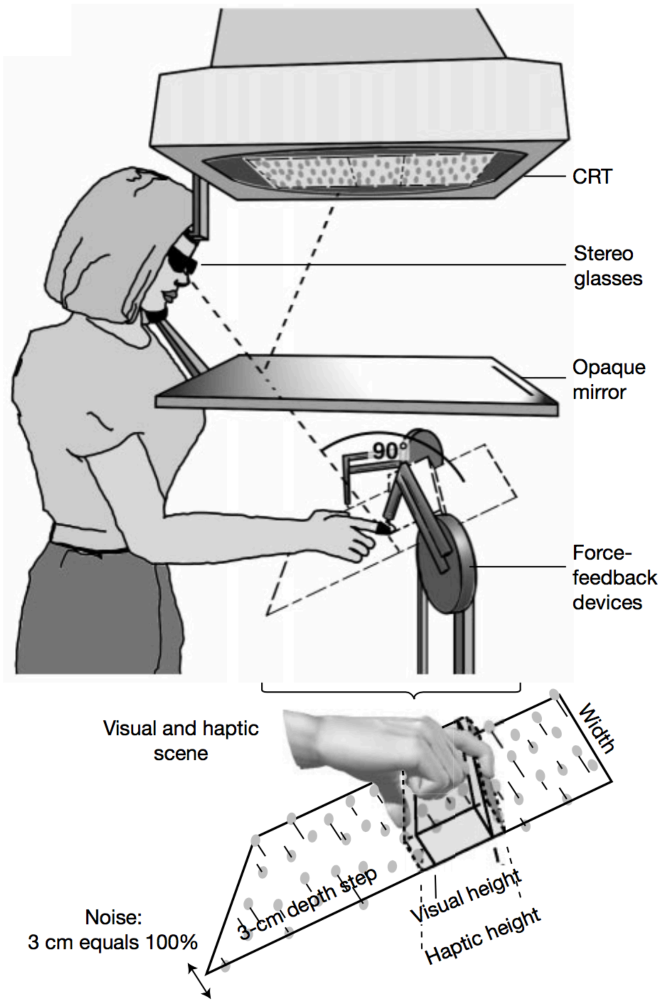


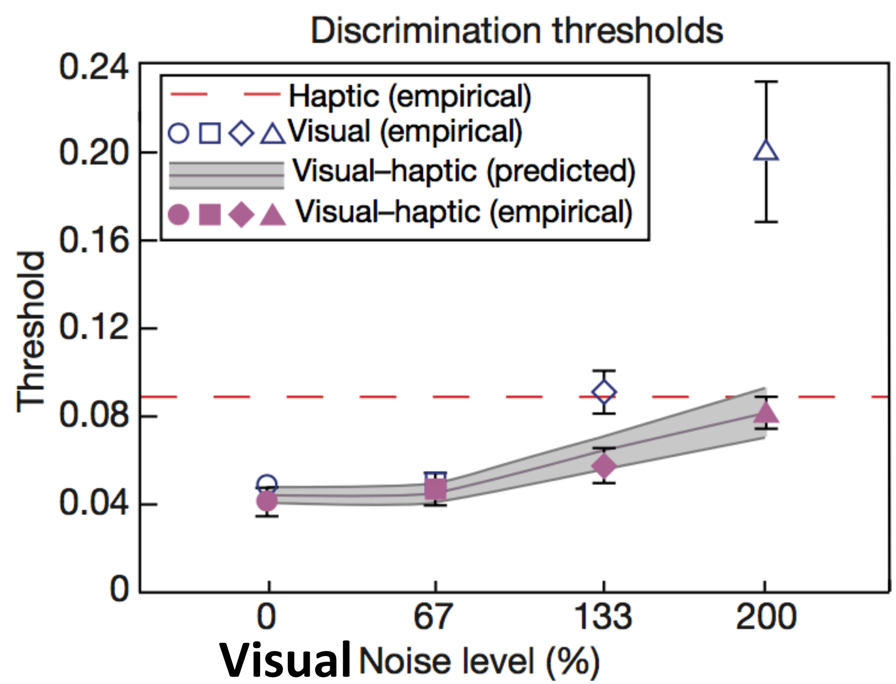
Varying offsets of visual to auditory information

Varying accuracy of visual information.

Alais & Burr, The **ventriloquist** effect results from near-optimal bimodal integration, Curr. Biol., 2004

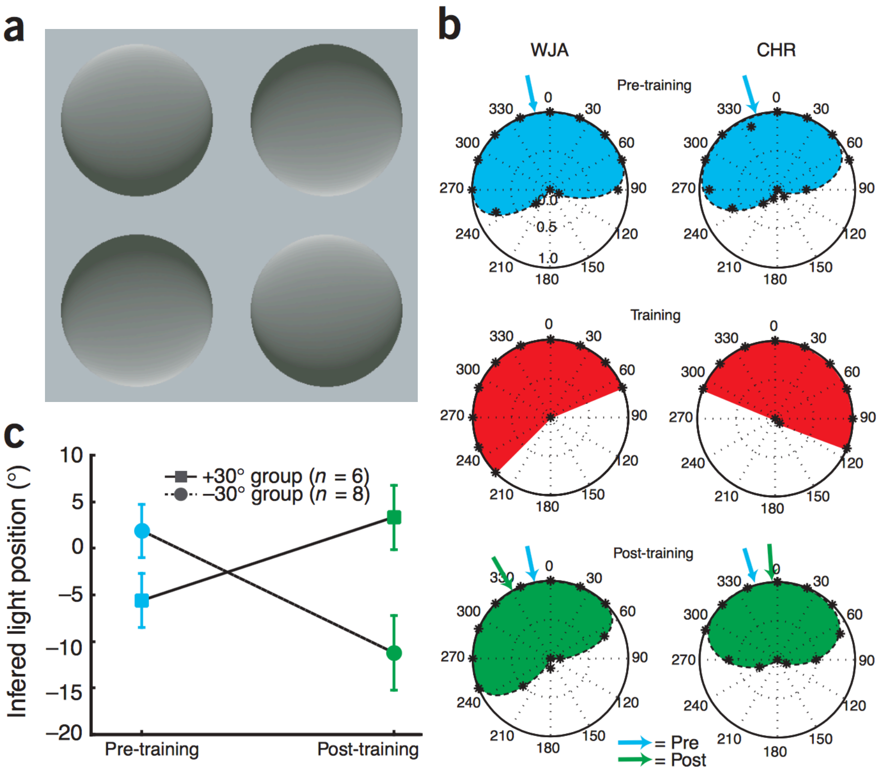
Sounds reasonable, but might it be true?





Ernst& Banks, Humans integrate visual and haptic information in a statistically optimal fashion, Nature, 2002

Sounds reasonable, but might it be true?



Adams, Graf & Ernst, Experience can change the 'light-from-above' prior, Nat. Neuroscience, 2004

The success story of Bayesian Models for Perception

[Friston and Stephan, 2007; Knill and Pouget, 2004; Knill and Richards, 1996].

Magnitude Estimation [Shadlen, Kiani, Glasauer, Petzschner ...]

Visual perception [Weiss, Simoncelli, Adelson, Richards, Freeman, Feldman, Kersten, Knill, Maloney, Olshausen, Jacobs, Pouget, ...]

Language acquisition and processing [Brent, de Marken, Niyogi, Klein, Manning, Jurafsky, Keller, Levy, Hale, Johnson, Griffiths, Perfors, Tenenbaum, …]

Motor learning and motor control [Ghahramani, Jordan, Wolpert, Kording, Kawato, Doya, Todorov, Shadmehr, …]

Associative learning [Dayan, Daw, Kakade, Courville, Touretzky, Kruschke, …]

Memory [Anderson, Schooler, Shiffrin, Steyvers, Griffiths, McClelland, …]

Attention [Mozer, Huber, Torralba, Oliva, Geisler, Yu, Itti, Baldi, …]

Categorization and concept learning [Anderson, Nosfosky, Rehder, Navarro, Griffiths, Feldman, Tenenbaum, Rosseel, Goodman, Kemp, Mansinghka, …]

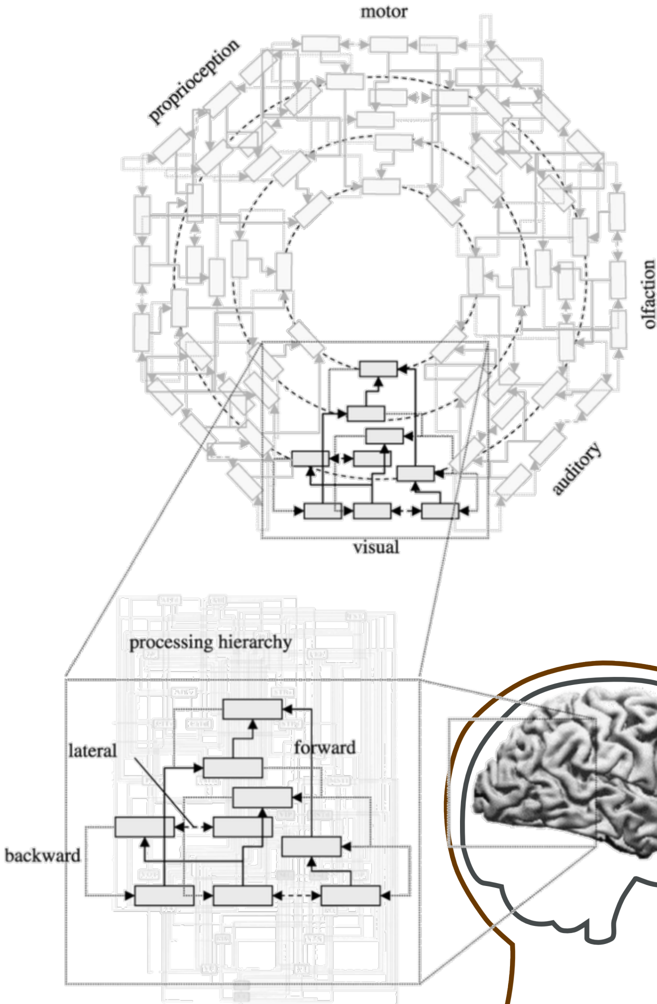
Reasoning [Chater, Oaksford, Sloman, McKenzie, Heit, Tenenbaum, Kemp, …]

Causal inference [Waldmann, Sloman, Steyvers, Griffiths, Tenenbaum, Yuille, …]

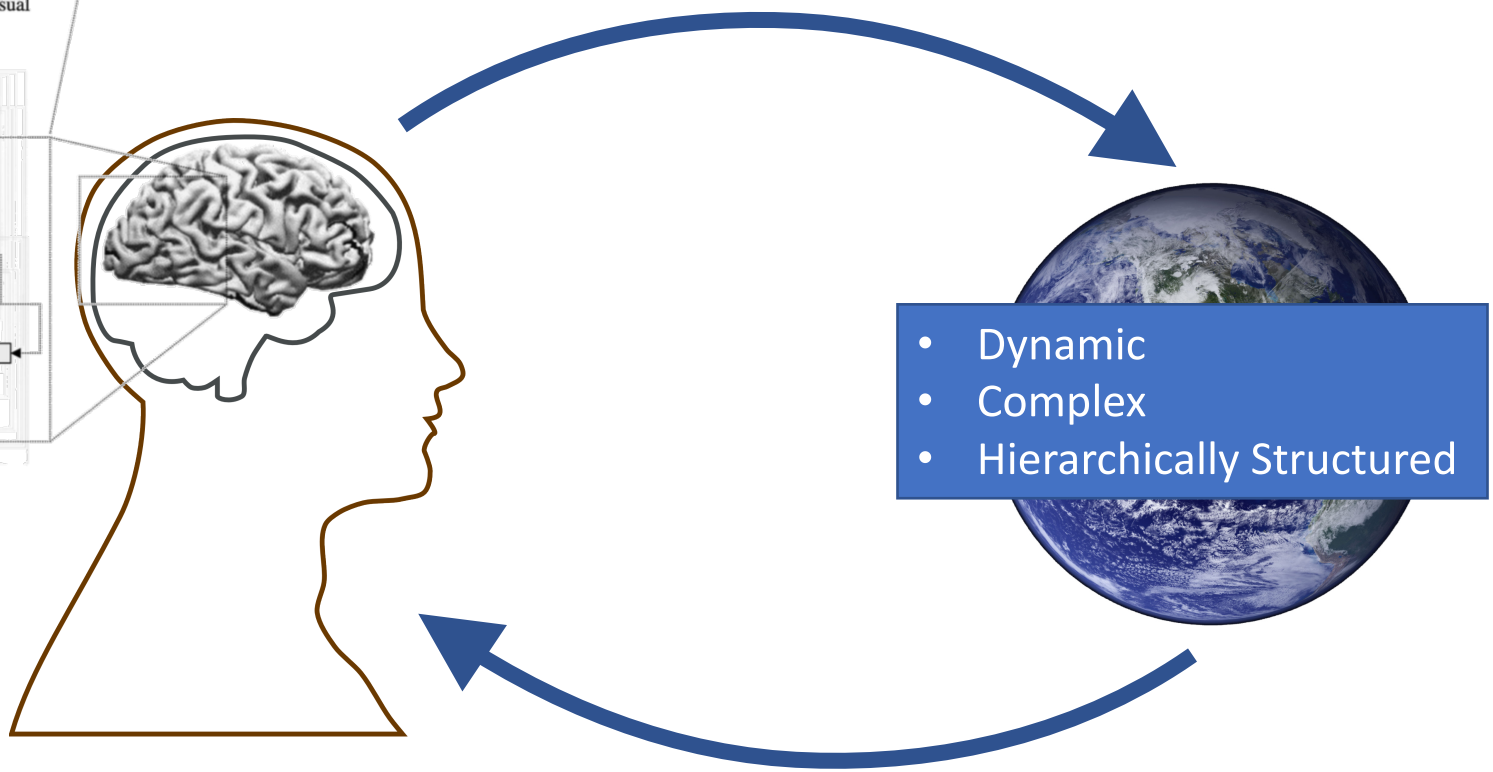
Decision making and theory of mind [Lee, Stankiewicz, Rao, Baker, Goodman, Tenenbaum, …]

F. Petzschner,
https://bitbucket.org/fpetzschner/cpc2016

How might Bayesian Inference be implemented in the Brain?*



Friston,
Phil. Trans.
R. Soc. B,
2005



*Disclaimer: Now it gets speculative!

Some Assumptions about Model Structure

Generative Model:

“Prior”: Pink elephants are not very common.

$$p(o(t), x(t)) = p(o(t)|x(t))p(x(t))$$

Observations:

“Likelihood”: How would a pink elephant look like?

- Vision: “A large pink thing in the shape of an elephant”
- Hearing: “Trooeeeet”
- Touch: The ground is vibrating

“A pink elephant is just right in front of me.”

Some Assumptions about Model Structure

Hidden Variables: $x = \{\theta, s(t)\}$

"Parameters", encode slowly changing dependencies, physical laws, general rules

"States", encode hidden reasons for observations on fast timescale, object identities, positions, physical properties, …

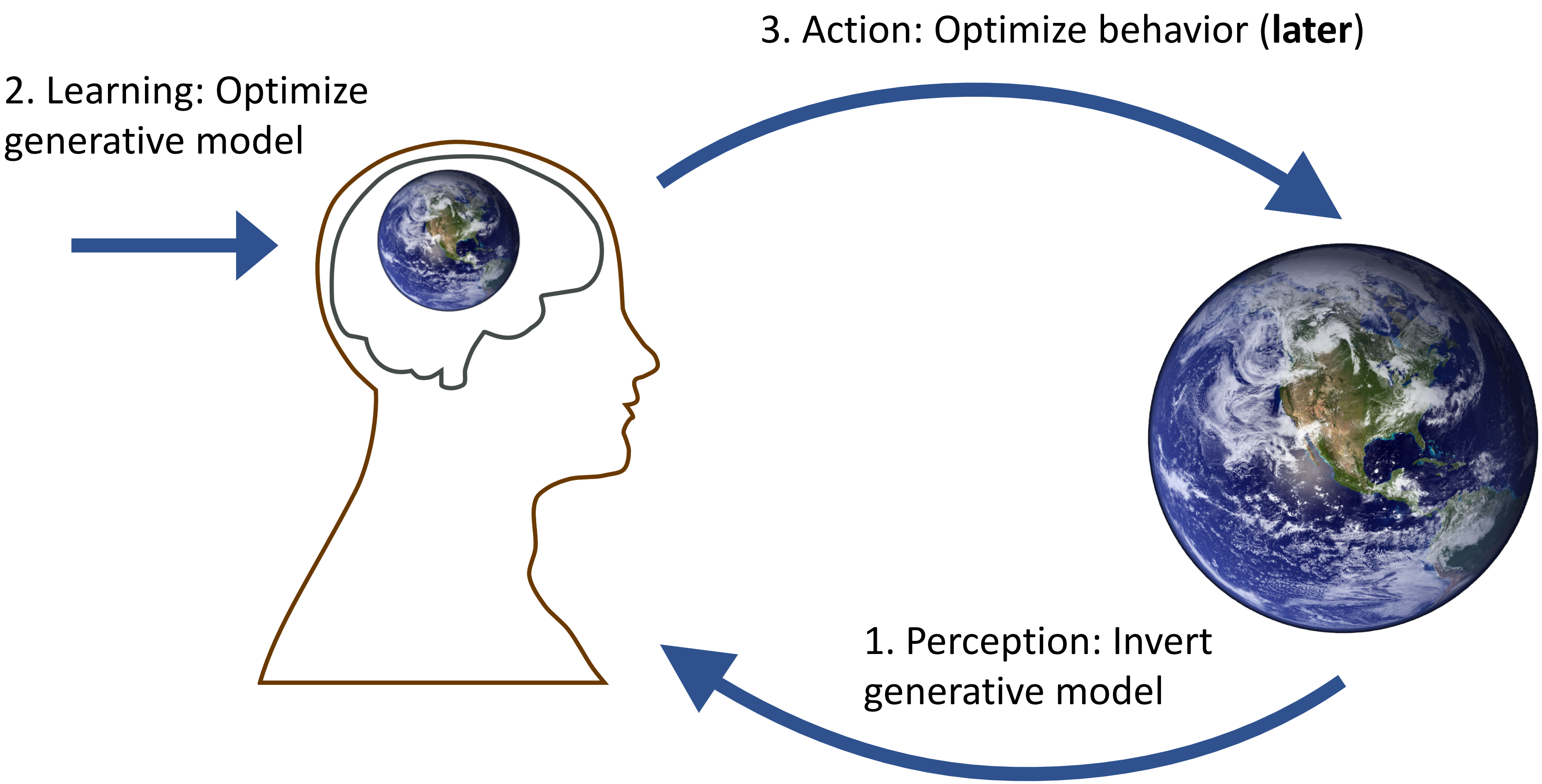
Hierarchy: $p(\theta, s(t)) = p(s(t)|\theta)p(\theta)$

The parameters (general laws) govern how the hidden states of the world (**which might have another hierarchy by themselves**) evolve

Factorization: $p(o(t)|\theta, s(t' \leq t)) = p(o(t)|\theta, s(t))$

My **sensory input right now** only depends on the general laws of the world and the **state of the world right now**.

Three very hard problems:



Problem 1: Perception (Inference on States)

Invert Generative Model using Bayes' Theorem:

"Likelihood": How would a pink elephant look like?

"Prior": Pink elephants are not very common.

$$p\big(s(t)|o(t)\big) = \frac{p(o(t)|s(t))p(s(t))}{p(o(t))}$$

"Maybe there is really a pink elephant right in front of me."

Observations: Vision: "A large pink thing in the shape of an elephant" Hearing: A loud trumpet. Touch: The ground is vibrating

It's not very likely, to make such observations.

Buuuuut:

$$p(o(t)|s(t)) = \int p(o(t)|s(t),\theta)p(\theta)\mathrm{d}\theta$$

$$p\big(o(t)\big) = \iint p(o(t)|s(t),\theta)p(s(t)|\theta)p(\theta)\mathrm{d}s(t)\mathrm{d}\theta$$

$$p\big(s(t)\big) = \int p(s(t)|\theta)p(\theta)\, \mathrm{d}\theta$$

Extremely high-dimensional integrals! Not even highly parallel computational architectures, such as the brain, can solve these **exactly**.

Problem 2: Learning (Inference on Parameters)

Given some observations $o(t_1), \dots, o(t_n)$ at times $t_1 < t_2 < \cdots < t_n$ use Bayes' Theorem to update parameters θ:

$$p(\theta | o(t_1), \dots, o(t_n)) = \frac{p(o(t_1),\dots,o(t_n)|\theta)p(\theta)}{p(o(t_1),\dots,o(t_n))}$$

"Now that I've seen a pink elephant, maybe they are not that unlikely after all…"

In „real time" the agent could update its parameters in the following way:

$$p(\theta | o(t_1), \dots, o(t_n)) = \frac{p(o(t_n)|\theta, o(t_1),\dots,o(t_{n-1}))\, p(\theta | o(t_1),\dots,o(t_{n-1}))}{p(o(t_n))}$$

This leads to comparatively „slow" update dynamics, compared to the dynamics of the hidden states, which might completely change according to the current observation.

Buuuuuut (again):

$$p(o(t_1), \dots, o(t_n)|\theta) = \int p(o(t_1), \dots, o(t_n), s(t_1), \dots, s(t_n)|\theta)\mathrm{d}s(t_1) \dots \mathrm{d}s(t_n)$$

$$p(o(t_1), \dots, o(t_n)) = \iint p(o(t_1), \dots, o(t_n), s(t_1), \dots, s(t_n), \theta)\mathrm{d}s(t_1) \dots \mathrm{d}s(t_n)\, \mathrm{d}\theta$$

Extremely high-dimensional integrals! Not even highly parallel computational architectures, such as the brain, can solve these.

Timescale of Perception

Given observations $o(t_1), \dots, o(t_n)$ at times $t_1 < t_2 < \dots < t_n$, the posterior probability on the state $s(t_n)$ at time t_n

$$p(s(t_n)|o(t_1), \dots, o(t_n)) = p(s(t_n)|o(t_n))$$

only depends on the current observation $o(t_n)$ at this time, and the time invariant parameters θ. I.e. as the state of the world changes very quickly (e.g. a tiger jumping into your field of view), the dynamics of the representation of the corresponding posterior distribution over states $s(t)$ are also very fast.

Timescale of Learning

As the agent makes observations $o(t_1), \dots, o(t_n)$ at times $t_1 < t_2 < \cdots < t_n$, the posterior probability on the parameters, given observations, gets a Bayesian update

$$p\big(\theta | o(t_1), \dots, o(t_n)\big) = \frac{p(o(t_n)|\theta, o(t_1), \dots, o(t_{n-1}))\, p(\theta|o(t_1), \dots, o(t_{n-1}))}{p(o(t_n))}$$

for each new observation, here shown for the last observation at t_n. The more observations the agent has made before, the more constrained its estimate $p(\theta|o(t_1), \dots, o(t_{n-1}))$ on the true parameters θ is already. I.e. while the representation of the posterior density on parameters, given observations, might initially change rather quickly, its dynamics will slow down the more the agent sees – and therefore learns – from its environment. Later on, strong evidence or many observations are required for large changes in the parameter estimates. Thus, the dynamics of the representation of the posterior density on the parameters will be rather slow.

Feynman, 1972

(A possible) solution: Variational Inference*

Recipe:

- Given observations $o = \{o(t_1), \dots, o(t_n)\}$, and generative model $p(o, \theta, s) = p(o|\theta, s)p(\theta, s)$, where s $= \{s(t_1), \dots, s(t_n)\}$
- Introduce approximation $q_\mu(\theta, s)$ to posterior density $p(\theta, s|o)$, parameterized by sufficient statistics $\mu = \{\mu_\theta, \mu_{s(t_1)}, \dots, \mu_{s(t_n)}\}$.

Converts a complex integration to an optimization problem.

Always ≥ 0, equal to 0 if and only if both distributions are equal. (But not symmetrical!)

- Minimize the variational free energy

$$F(o, \mu) = -\ln p(o) + \boxed{D_{\mathrm{KL}}\left(q_\mu(\theta, s) \,||\, p(\theta, s|o)\right)}$$

- This will maximize the evidence $p(o)$ for the agent's model of the world, while simultaneously **driving $\boldsymbol{q_\mu(\theta, s)}$ towards** the **true posterior $\boldsymbol{p(\theta, s|o)}$**.

***Disclaimer: Will be combined with Monte-Carlo Sampling next week! ☺**

Short interrupt: KL-Divergence

$$D_{\mathrm{KL}}\left(q_\mu(\theta,s)\,||\,p(\theta,s|o)\right) = \left\langle \ln \frac{q_\mu(\theta,s)}{p(\theta,s|o)} \right\rangle_{q_\mu(\theta,s)}$$

Expectation with respect to $q_\mu(\theta,s)$

It's really easy to evaluate for Gaussians:

$$D_{\mathrm{KL}}\left(N(x;\mu_1,\sigma_1)\,||\,N(x;\mu_2,\sigma_2)\right) = \left\langle \ln \frac{N(x;\mu_1,\sigma_1)}{N(x;\mu_2,\sigma_2)} \right\rangle_{N(x;\mu_1,\sigma_1)}$$

$$= \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

What have we won?

To minimize, we have to **evaluate** the variational Free Energy

$$F(o,\mu) = -\ln p(o) + D_{\mathrm{KL}}\left(q_\mu(\theta,s)\,||\,p(\theta,s|o)\right)$$

How hard to evaluate: 🐍* 🐶 🐉

This is just the posterior, that we want to approximate!

can be rewritten as

$$F(o,\mu) = < -\ln p(o|\theta,s) >_{q_\mu(\theta,s)} + D_{\mathrm{KL}}\left(q_\mu(\theta,s)\,||\,p(\theta,s)\right)$$

"Complexity"

How hard to evaluate: 🐹 "Accuracy" 🐶 🐣

or (for Physicists):

$$F(o,\mu) = < -\ln p(o,\theta,s) >_{q_\mu(\theta,s)} - < -\ln q_\mu(\theta,s) >_{q_\mu(\theta,s)}$$

How hard to evaluate: 🐨 Expected Energy 🐶 Entropy

*illustration of variational inference with emojis from: http://www.inference.vc/choice-of-recognition-models-in-vaes-a-regularisation-view/

Predictive Coding

Assume simplest way of minimizing F possible:
Gradient Descent

The sufficient statistics μ_θ and μ_s change to minimize the Free Energy $\mathrm{F}(\mu_\theta, \mu_s, o)$ via:

$$\dot{\mu_\theta} \propto -\nabla_{\mu_\theta}\mathrm{F}(\mu_\theta, \mu_s, o)$$
$$\dot{\mu_s} \propto -\nabla_{\mu_s}\mathrm{F}(\mu_\theta, \mu_s, o)$$

The dynamics of the **sufficient statistics $\boldsymbol{\mu_\theta}$** of the approximate posterior density over **parameters $\boldsymbol{\theta}$ of the generative model** are very slow:
→ $\boldsymbol{\mu_\theta}$ can be represented in terms of **synaptic connectivity**.

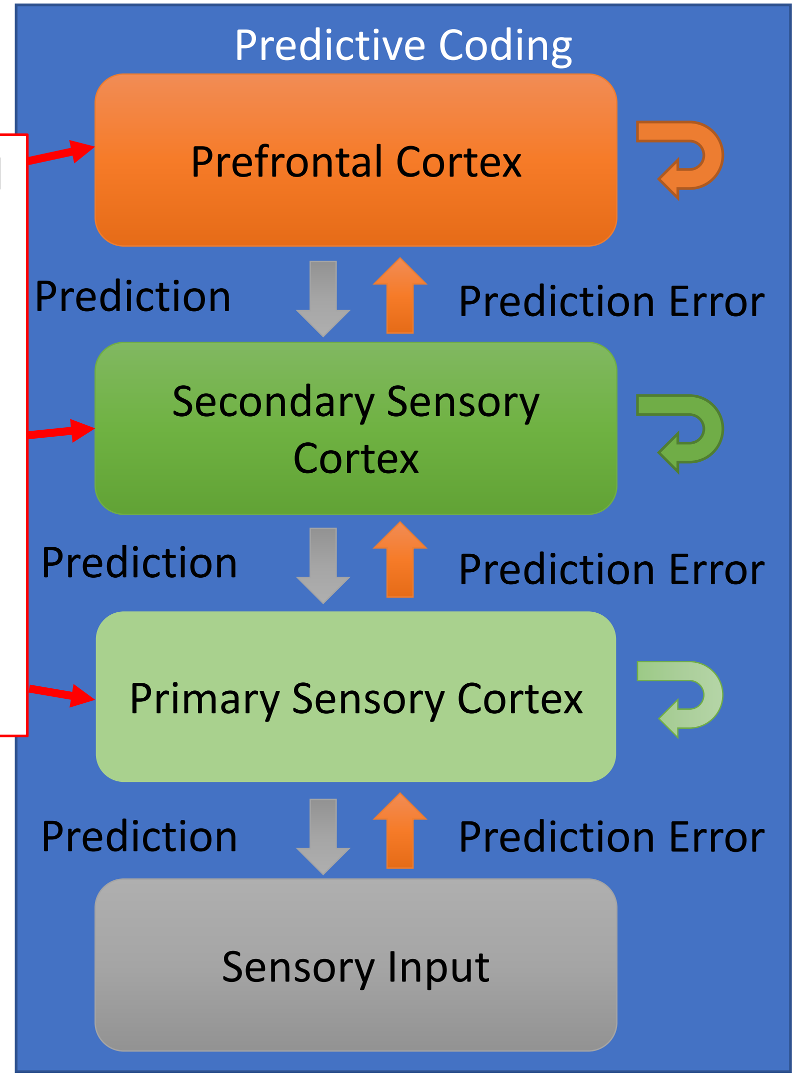
The dynamics of the **sufficient statistics $\boldsymbol{\mu_s}$** of the approximate posterior density over **hidden states $\boldsymbol{s}$** are fast:
→ $\boldsymbol{\mu_s}$ can be represented in terms of **neural activity**.

Predictive Coding:

Additional assumptions about the structure and implementation of the states $s(t)$:

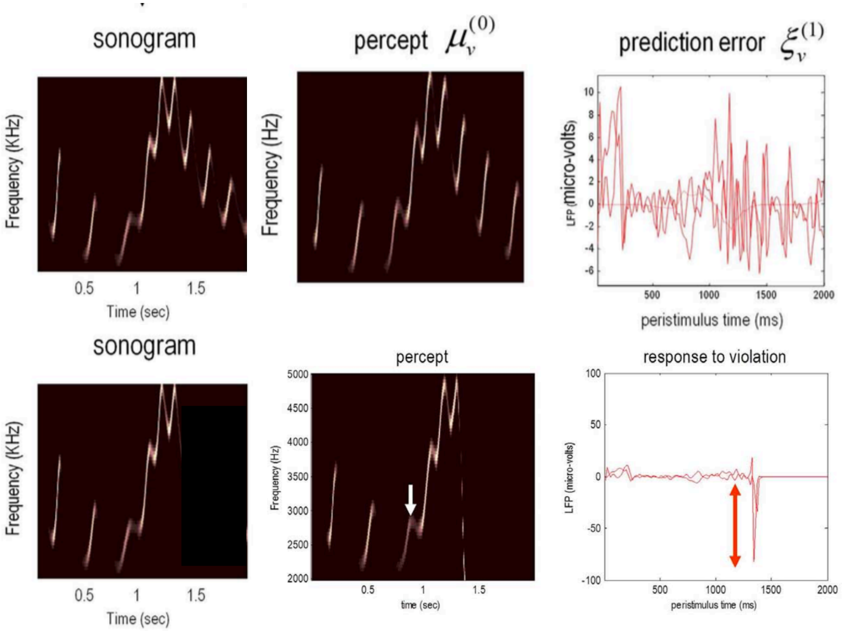
- Probabilities represented by Gaussians, where sufficient statistics $\mu_s(t)$ and μ_θ represent means **and covariance** matrices. **Inverse covariance matrix** = **“precision”**
- Hierarchical temporal structure of states $s(t)$.

Recurrent neural dynamics at each level can implement attractor networks, winner-take-all networks, winnerless competition,…

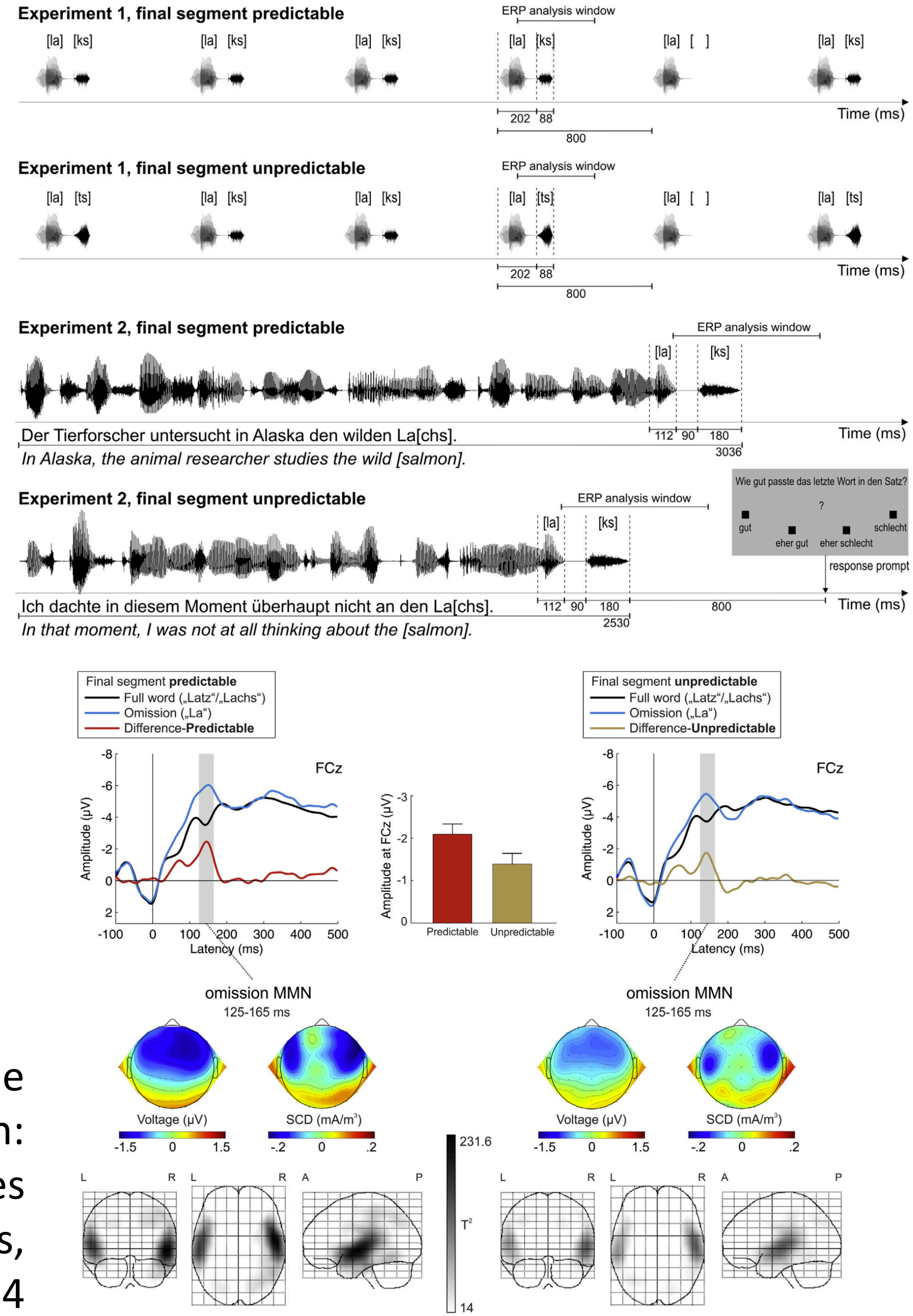


c.f. Friston,
Phil. Trans. R. Soc. B, 2005

Reality check

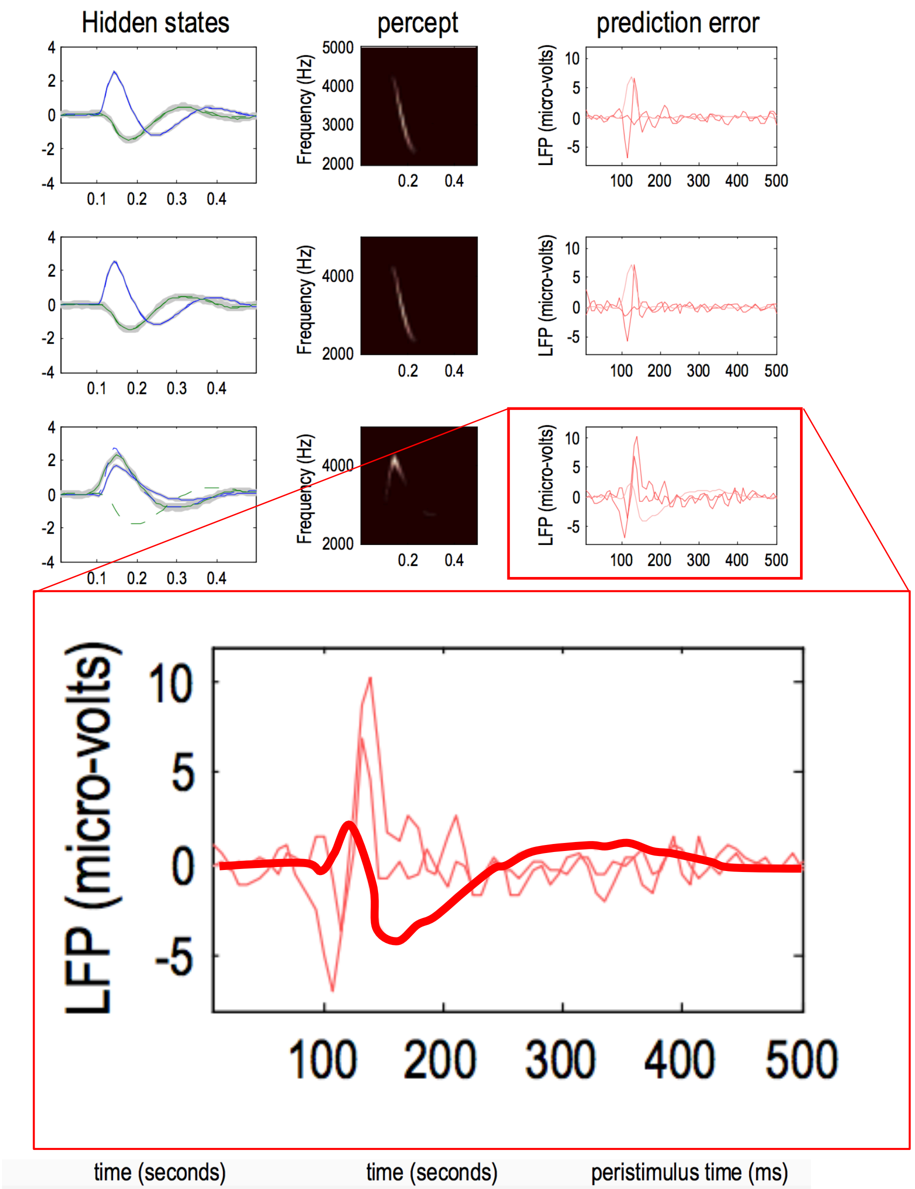


Adams et al., The Computational Anatomy of Psychosis, Front. Psychiatry, 2014

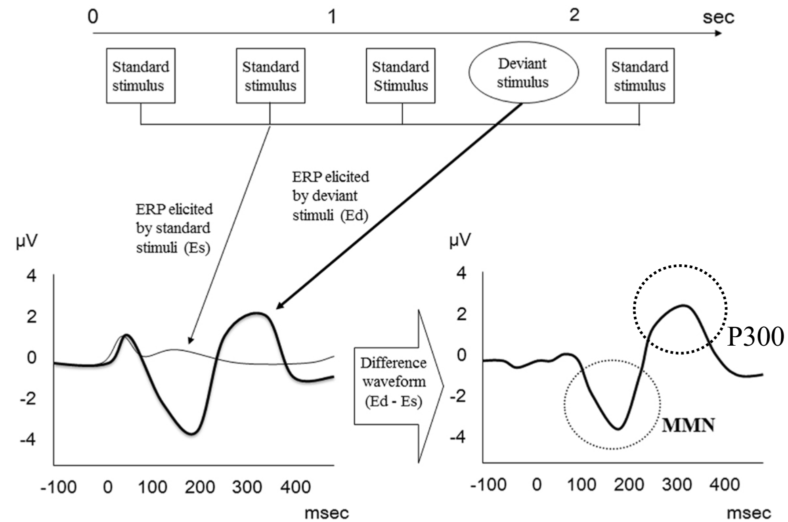


Bendixen et al., Prediction in the service of comprehension: Modulated early brain responses to omitted speech segments, Cortex, 2014

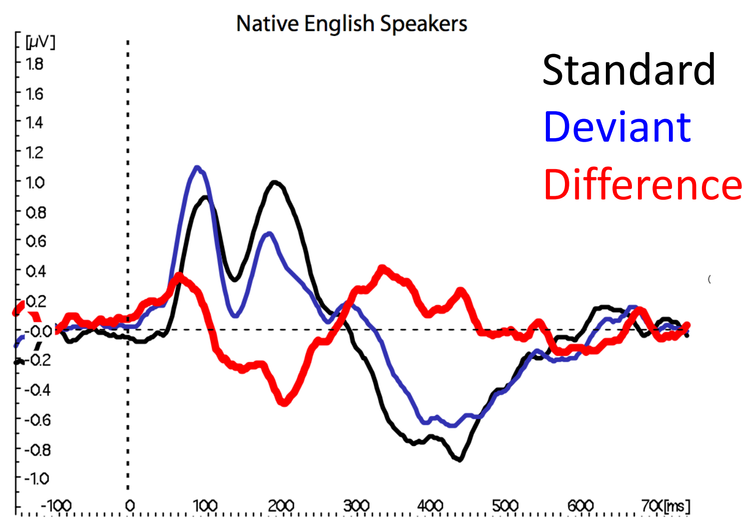
Reality check



Friston & Kiebel, Attractors in Song, New Mathematics and Natural Computation, 2009,



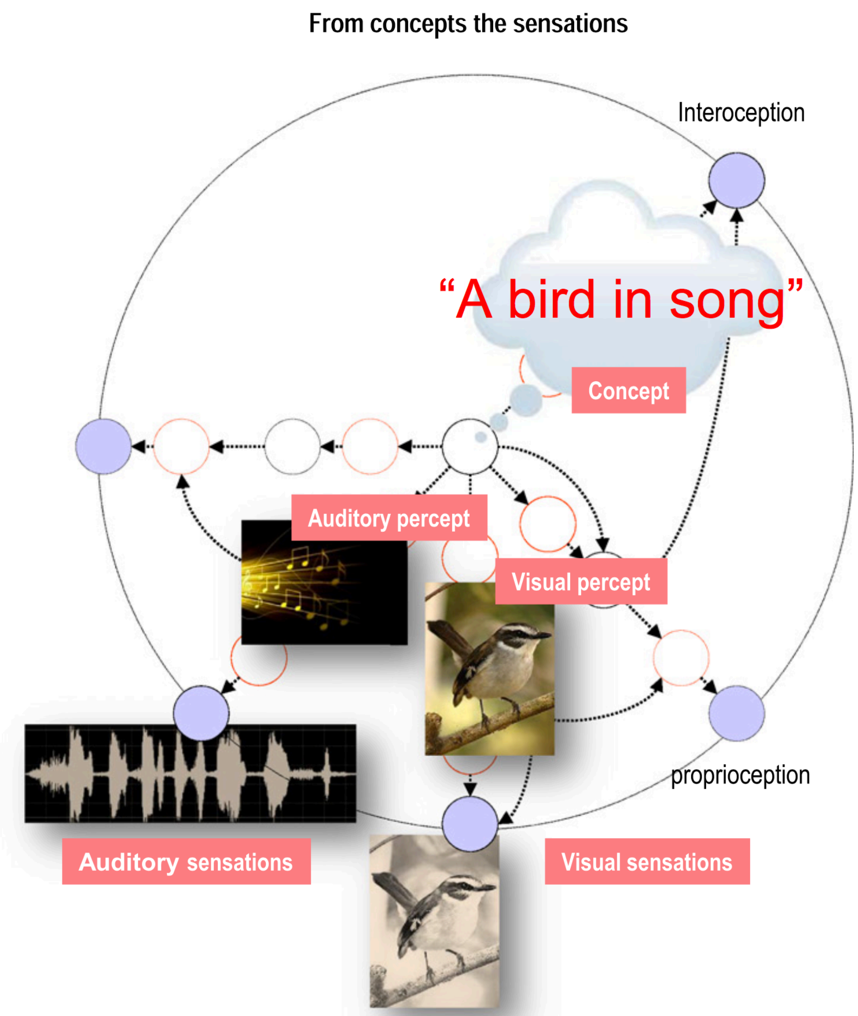
Nagai et al., Front. Psychiatry, 2013



Zevin et al., Front. Hum. Neurosci., 2010

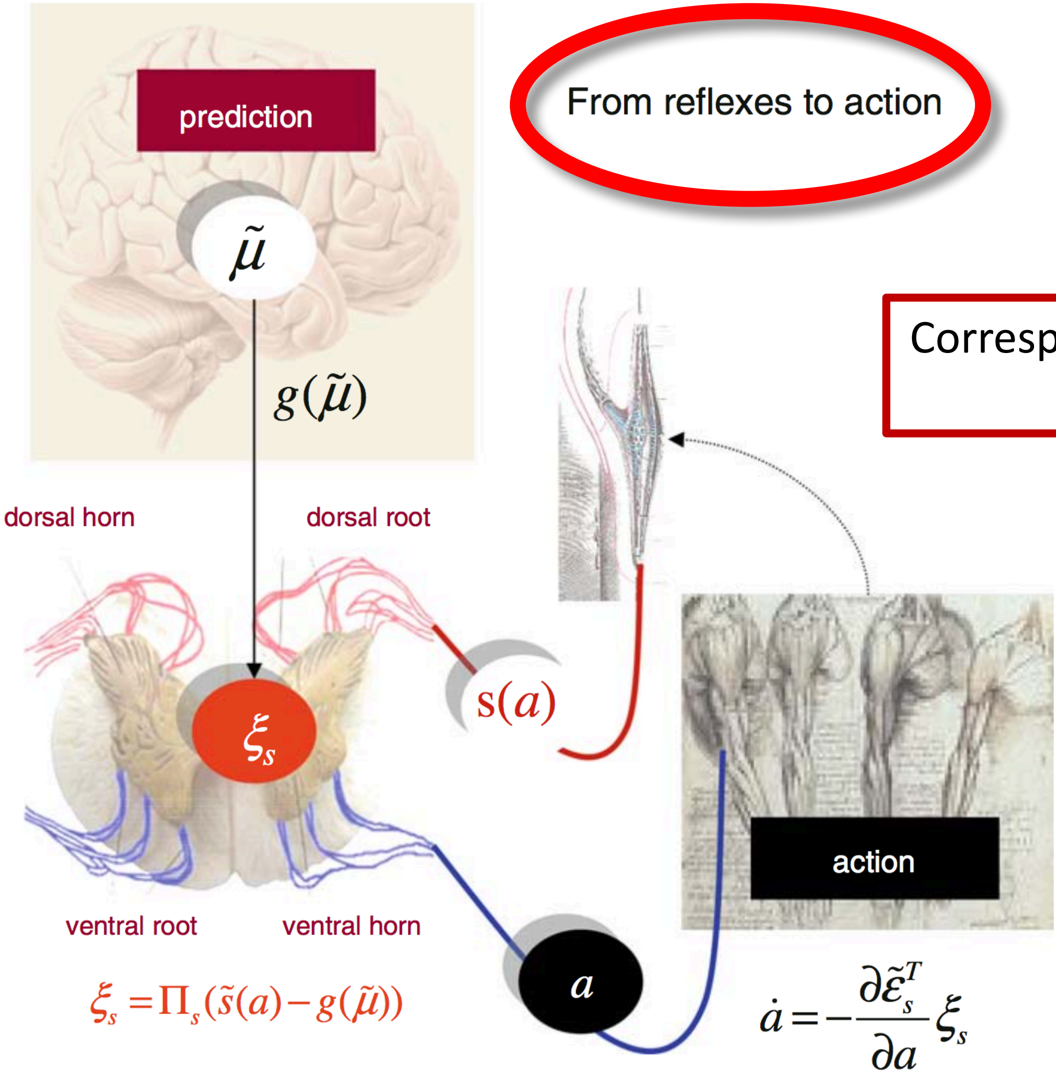
Predictive Coding Summary

- Our brain uses a variational approximation to invert and optimize a generative model of its sensations.
- The model corresponds to the world, i.e. it is nonlinear, dynamic and hierarchically structured.
- The posterior on states is represented by means of neural activity, the posterior on parameters is represented by means of synaptic connectivity.
- Using simple assumptions about the hierarchical form, the distributions (Gaussians) and the optimization (Gradient Descent), the resulting predictive coding scheme matches cortical hierarchies, behavioral data, and neurophysiological responses, such as repetition suppression, omission responses, and mismatch negativity.



Bastos et al., Canonical Microcircuits for Predictive Coding, Neuron, 2012

Active Inference: Predictive Coding with Reflex Arcs



From reflexes to action

Corresponds nicely to the architecture of our motor system.

Friston, Daunizeau, Kilner, Kiebel, Action and behavior: a free-energy formulation, Biol. Cybernetics, 2011

How to formulate this?

Remember the following form of the variational Free Energy:

$$F(o,\mu) =< -\ln p(o|\theta,s) >_{q_\mu(\theta,s)} + D_{\mathrm{KL}}\big(q_\mu(\theta,s)\,||\,p(\theta,s)\big)$$

How hard to evaluate: 🐹 "Accuracy" 🐶 🐣 "Complexity"

The **accuracy** term depends on **observations**, which in turn depend on the current, true **state of the world**, which again depends on the agent's **actions**.

By choosing **actions** $\boldsymbol{a(t)}$, in terms of the states of output organs (muscles, mainly…), to **minimize variational Free Energy**, the agent will **seek out sensations**, that are likely under its generative model of the world and its current beliefs about the state of the world.

Summary: Active Inference

The sufficient statistics

- μ_θ of the parameters of the generative model
- μ_s of the hidden states of the world
- μ_a of the states of the agent's effector organs

all change to minimize the variational Free Energy F

$$(\mu_\theta, \mu_s, \mu_a) = \underset{\mu_\theta^*, \mu_s^*, \mu_a^*}{\operatorname{argmin}} F(o(\mu_a^*), \mu_\theta^*, \mu_s^*)$$

Where

$$F(o, \mu) = < -\ln p(o|\theta, s) >_{q_\mu(\theta,s)} + D_{\mathrm{KL}}\big(q_\mu(\theta, s) \,||\, p(\theta, s)\big)$$

Some preliminary thoughts…

- Right now Active Inference gives an abstract account of the hierarchical architecture of the cortex, the basic architecture of the motor system, perceptual phenomena, and macroscopic neural responses.
- But we used a looooooong list of assumptions and seemingly counter-intuitive arguments, i.e.

Does this view of action not implicate, that I should retire to a dark room and turn off the light? I would be able to exactly predict my sensory input and all would be fine. Well, …

at some point, you would get **thirsty.**

Evolutionary Argument

- In order to survive, an agent has to keep certain inner parameters within very strict bounds.
- Thus, it has to constrain the entropy of the probability distributions over these parameters.
- But entropy is just:

$$H(S) = < -\ln p(s) >_{p(s)}$$

- Assuming we have sensory systems, that give us access to the relevant parameters (glomus caroticus, osmoreceptors in the hypothalamus, macula densa, …) this can be upper bounded by:

$$H(S) \leq H(O) + \text{const.}$$

- Where

$$H(O) = < -\ln p(o) >_{p(o)} = \lim_{T \to \infty} \frac{1}{T} \int_0^T -\ln p\big(o(t)\big) \mathrm{d}t$$

Ergodicity

$$\boxed{\text{The agent can keep its physiological variables within viable bounds by minimizing sensory surprise at all times (Euler-Lagrange-Equation).}}$$

Closing the circle…

Variational Free Energy is just:

$$F(o,\mu) = -\ln p(o) + \overbrace{\boxed{D_{\mathrm{KL}}\left(q_\mu(\theta,s)\,||\,p(\theta,s|o)\right)}}^{\geq 0}$$

How hard to evaluate: 🐍 🐶 🐉

- By minimizing Free Energy using action, an agent upper bounds its sensory surprise.
- Thereby, it can counteract dispersive effects of the environment, to sustain its physiological variables (e.g. its inner milieu) within viable bounds.
- So the Bayes-optimal **learning and perception** that we started with is only a by-product, required to **make** the **Free Energy**, which can be **evaluated** and **influenced** by the agent, a **tight bound** on **sensory surprise**, to allow for an agent's **survival**.

Closing the circle…

Variational Free Energy is just:

$$F(o,\mu) = -\ln p(o) + \boxed{D_{\mathrm{KL}}\left(q_\mu(\theta,s)\,||\,p(\theta,s|o)\right)} \geq 0$$

“Goals” or “Utility” in terms of prior expectations on states to be in, $p(\theta, s)$. States to be **highly frequented** are associated with **“high reward”**. → **Next Week**

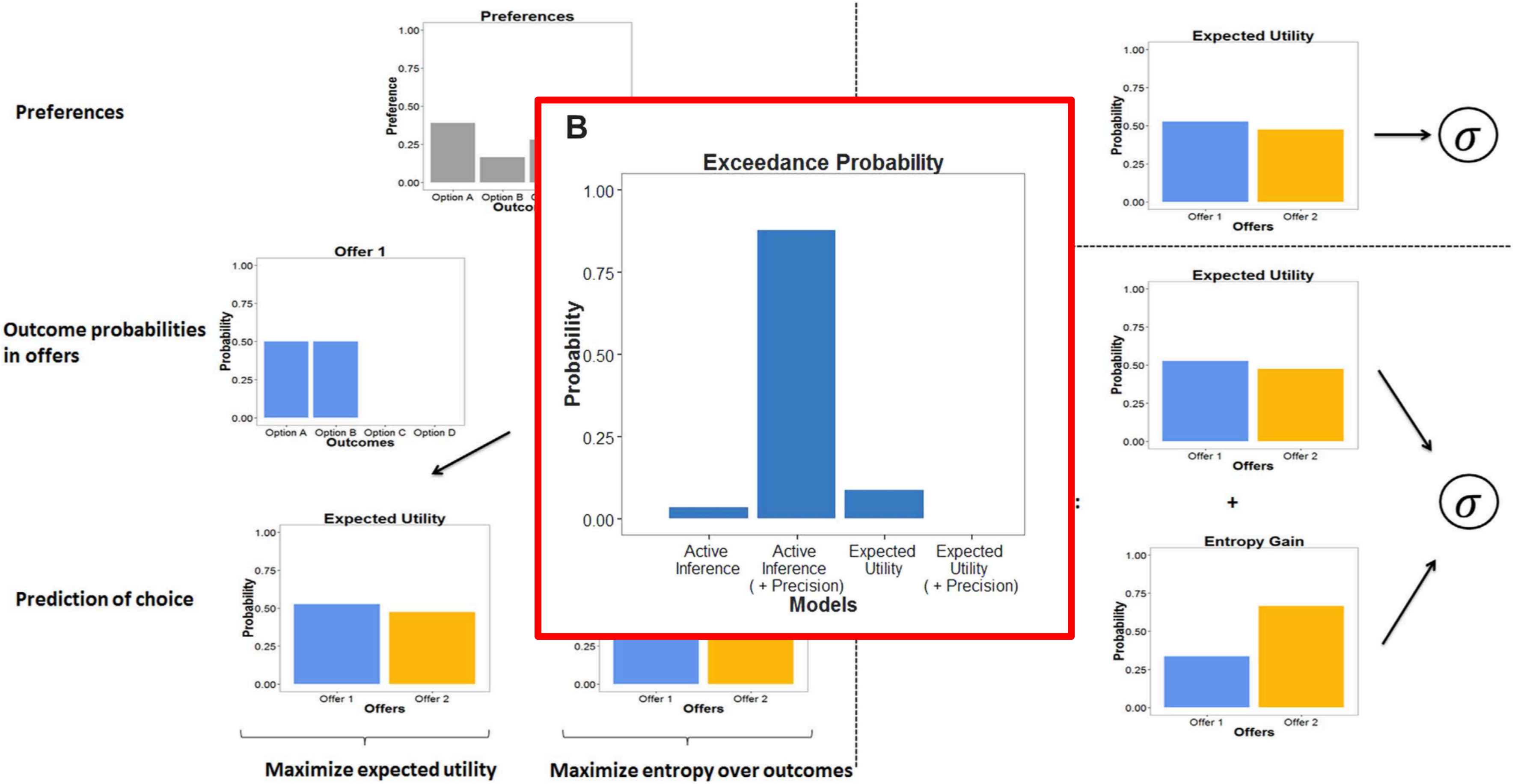
$$= < -\ln p(o|\theta,s) >_{q_\mu(\theta,s)} + \boxed{D_{\mathrm{KL}}\left(q_\mu(\theta,s)\,||\,p(\theta,s)\right)}$$

$$= < -\ln p(o,\theta,s) >_{q_\mu(\theta,s)} - < -\ln q_\mu(\theta,s) >_{q_\mu(\theta,s)}$$

$$= < -\ln p(o|\theta,s)\boxed{p(\theta,s)} >_{q_\mu(\theta,s)} - \boxed{< -\ln q_\mu(\theta,s) >_{q_\mu(\theta,s)}}$$

Maximize entropy of variational density → Keeping your options open, Novelty Seeking, Curiosity

Some First Evidence



Schwartenbeck et. al, Evidence for surprise minimization over value maximization in choice behavior, Scientific Reports, 2015