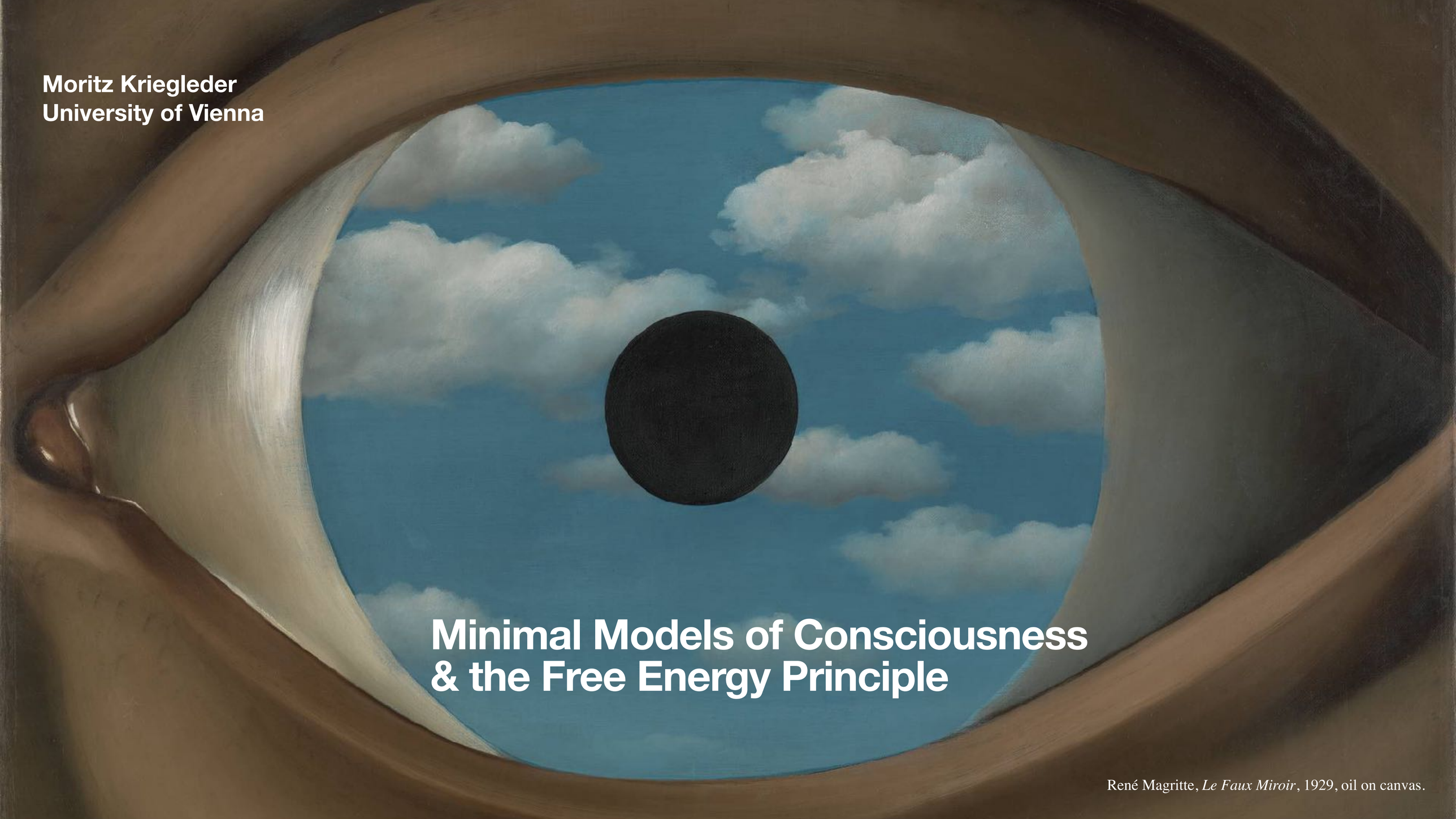
Moritz Kriegleder
University of Vienna

Minimal Models of Consciousness & the Free Energy Principle



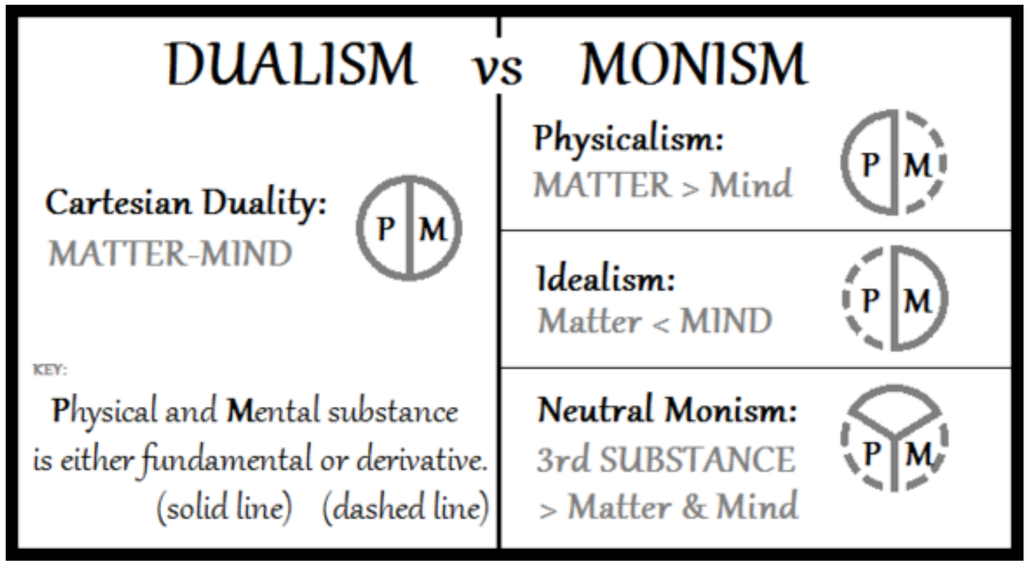
René Magritte, *Le Faux Miroir*, 1929, oil on canvas.

Overview

- So many models of consciousness
- We need a minimal model for unification and comparison
- The free energy principle as minimal model of consciousness?
- Difficulties of interpreting computational models
- Model templates as shared mathematical toolbox

Models of Consciousness

- We have a plethora of models available
- Divisible by either philosophical assumptions or mathematical tools
- From the quantum scale (Penrose, Fisher), over informational approaches (IIT, FEP) to high-level theories (Baars, Graziano)
- Dualism: Descartes
  Physicalism: Dennett
  Idealism: Berkeley
  Neutral Monism: Chalmers



Theory	Primary claim
Higher-order theory (HOT)	Consciousness depends on meta-representations of lower-order mental states
Self-organizing meta-representational theory	Consciousness is the brain's (meta-representational) theory about itself
Attended intermediate representation theory	Consciousness depends on the attentional amplification of intermediate-level representations
Global workspace theories (GWTs)	Consciousness depends on ignition and broadcast within a neuronal global workspace where fronto-parietal cortical regions play a central, hub-like role
Integrated information theory (IIT)	Consciousness is identical to the cause–effect structure of a physical substrate that specifies a maximum of irreducible integrated information
Information closure theory	Consciousness depends on non-trivial information closure with respect to an environment at particular coarse-grained scales
Dynamic core theory	Consciousness depends on a functional cluster of neural activity combining high levels of dynamical integration and differentiation
Neural Darwinism	Consciousness depends on re-entrant interactions reflecting a history of value-dependent learning events shaped by selectionist principles
Local recurrency	Consciousness depends on local recurrent or re-entrant cortical processing and promotes learning
Predictive processing	Perception depends on predictive inference of the causes of sensory signals; provides a framework for systematically mapping neural mechanisms to aspects of consciousness
Neuro-representationalism	Consciousness depends on multilevel neurally encoded predictive representations
Active inference	Although views vary, in one version consciousness depends on temporally and counterfactually deep inference about self-generated actions
Beast machine theory	Consciousness is grounded in allostatic control-oriented predictive inference
Neural subjective frame	Consciousness depends on neural maps of the bodily state providing a first-person perspective
Self comes to mind theory	Consciousness depends on interactions between homeostatic routines and multilevel interoceptive maps, with affect and feeling at the core
Attention schema theory	Consciousness depends on a neurally encoded model of the control of attention
Multiple drafts model	Consciousness depends on multiple (potentially inconsistent) representations rather than a single, unified representation that is available to a central system
Sensorimotor theory	Consciousness depends on mastery of the laws governing sensorimotor contingencies
Unlimited associative learning	Consciousness depends on a form of learning which enables an organism to link motivational value with stimuli or actions that are novel, compound and non-reflex inducing
Dendritic integration theory	Consciousness depends on integration of top-down and bottom-up signalling at a cellular level
Electromagnetic field theory	Consciousness is identical to physically integrated, and causally active, information encoded in the brain's global electromagnetic field
Orchestrated objective reduction	Consciousness depends on quantum computations within microtubules inside neurons

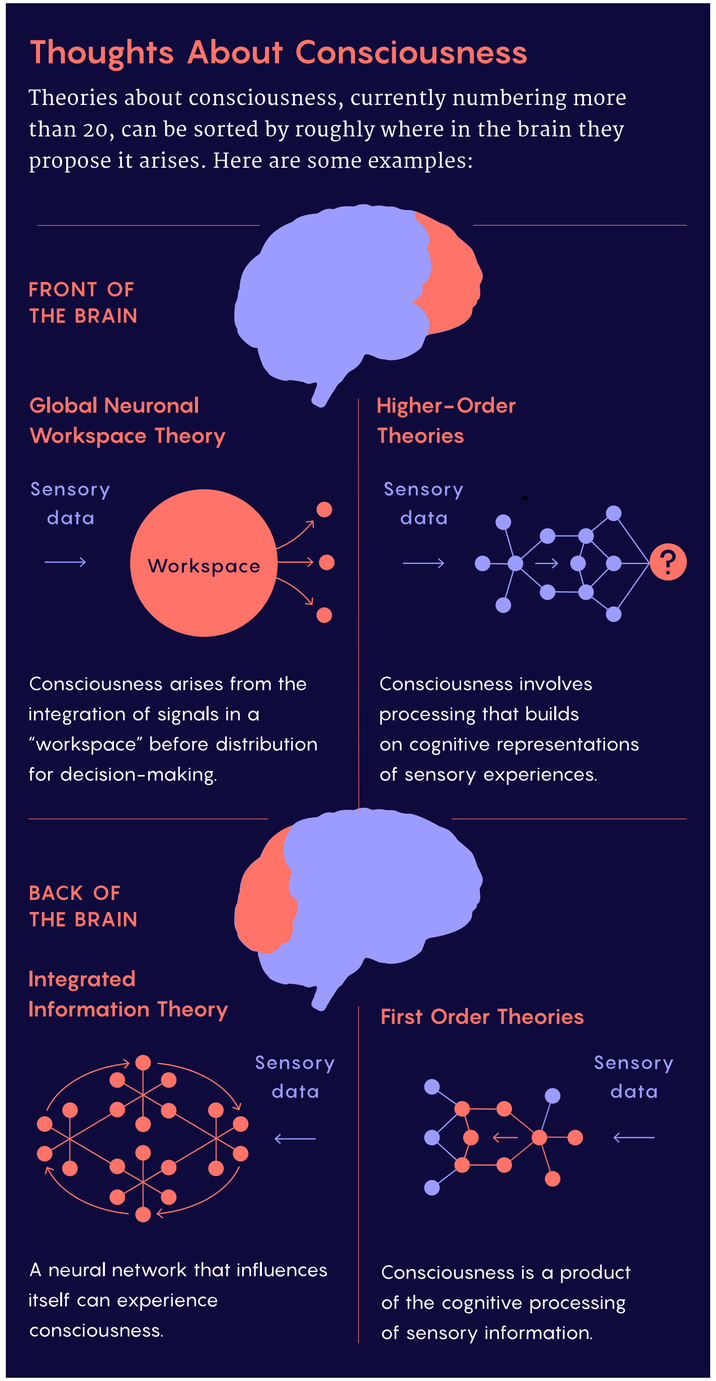
From: Seth & Bayne 2022

Models of Consciousness

- My project focuses on mathematical models of phenomenal consciousness
- Lots of research into specific models but little about comparison and unification but no common ground!
- Transfer of mathematical tools and concepts could help to track ideas
- *"An organism has conscious mental states if and only if there is something that it is like to be that organism – something that it is like for the organism."*
  Nagel 2005 p. 637

Neuroscience of Consciousness

- Neural Correlates of Consciousness (Crick & Koch 1990)
- *"Phenomenal consciousness is experience; the phenomenally conscious aspect of a state is what it is like to be in that state. The mark of access-consciousness, by contrast, is availability for use in reasoning and rationally guiding speech and action."* Block, 1997 p. 227
- Goal to find maps between phenomenal consciousness & brain activity
- Tests in adversarial collaborations forthcoming



From: https://www.quantamagazine.org/what-a-contest-of-consciousness-theories-really-proved-20230824/

Free Energy Models 1

- Variational free energy F is an upper bound on how surprising environmental states are
- It quantifies the difference between prediction and perception -> prediction error minimisation
- Belief updating satisfies Bayes rule

$$
\begin{aligned}
F[Q,y] &= \underbrace{-\mathbb{E}_{Q(x)}[\ln P(y,x)]}_{Energy} - \underbrace{H[Q(x)]}_{Entropy} \\
&= \underbrace{D_{KL}[Q(x) \,||\, P(x)]}_{Complexity} - \underbrace{\mathbb{E}_{Q(x)}[\ln P(y|x)]}_{Accuracy} \\
&= \underbrace{D_{KL}[Q(x) \,||\, P(x|y)]}_{Divergence} - \underbrace{\ln P(y)}_{Evidence}
\end{aligned}
$$

- **Example:** Dark room problem

Free Energy Models 2

- Expected free energy G predicts outcomes of different actions
- It is used to choose the best policy π balancing exploration and exploitation
- Dark room has low immediate surprise but provides no information of future survival

$$G(\pi) = -\underbrace{\mathbb{E}_{Q(\tilde{x},\tilde{y}|\pi)}[D_{KL}[Q(\tilde{x}|\tilde{y},\pi)\,||\,Q(\tilde{x}|\pi)]]}_{\text{Information gain}} - \underbrace{\mathbb{E}_{Q(\tilde{y}|\pi)}[\ln P(\tilde{y}|C)]}_{\text{Pragmatic value}}$$

$$= \underbrace{\mathbb{E}_{Q(\tilde{x}|\pi)}[H[P(\tilde{y}|\tilde{x})]]}_{\text{Expected ambiguity}} + \underbrace{D_{KL}[Q(\tilde{y}|\pi)\,||\,P(\tilde{y}|C)]}_{\text{Risk (outcomes)}}$$

Minimal models of consciousness

- Minimal unifying models (MUM) are necessary conditions for consciousness (Wiese 2020):

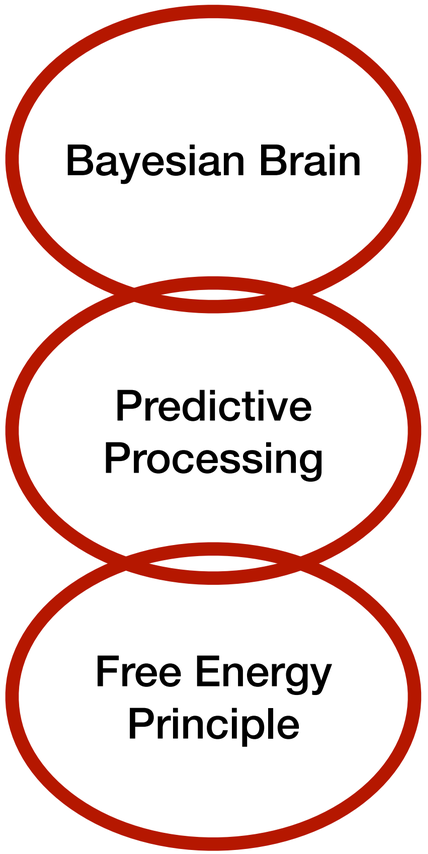
  1. Empirically minimal: specifies only necessary features
  2. Conceptually minimal: determinable characterisations of properties
  3. Minimally unifying: highlights common assumptions of different approaches

- Information generation seems necessary but not sufficient for consciousness

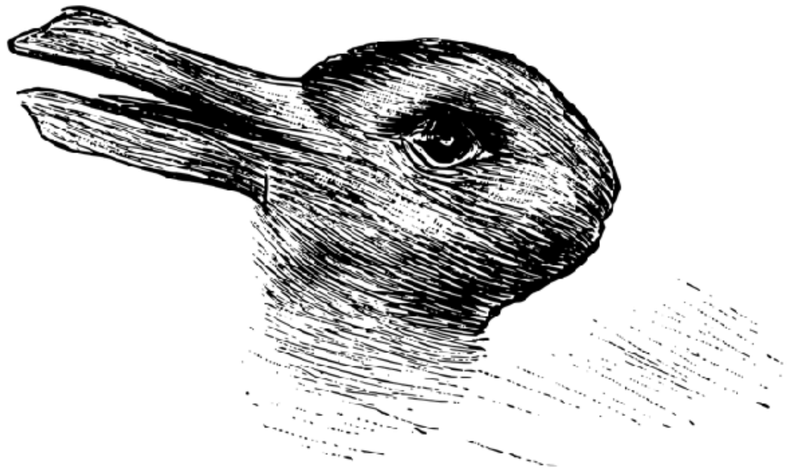
- “consciousness = information generation” is on its own empty,
  action & perception loops needed to account for many characteristics

Free Energy and Predictive Processing

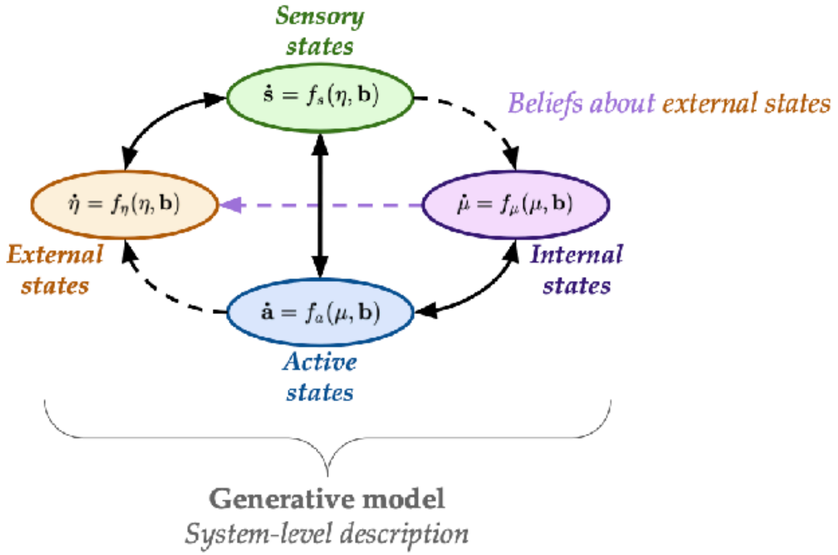
- Perception is not passive processing of external inputs but a comparison of predictions with external stimuli (Clark 2015)
- Minimising free energy is a cognitive agents strategy to minimise surprise about the environment (Friston 2009)
- FEP trades natural attitude - that we perceive things as they are - for active sense-making
- Has been used recently to explain phenomenal consciousness (Ramstead et al. 2022)



Rabbiduck Example

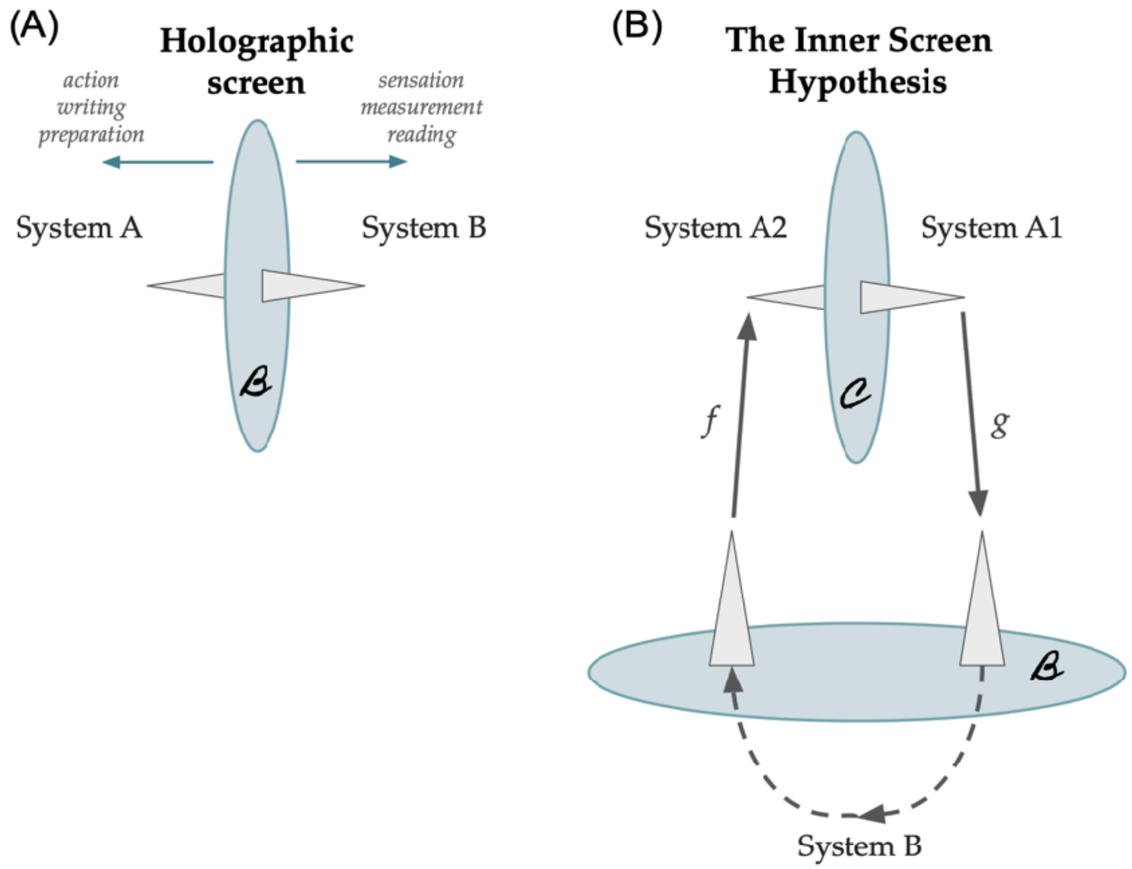


From: Bruegger & Brugger 1993



From: Ramstead & Albarracin 2023, preprint

Inner Screen Model of Consciousness



- Holographic screens separate predictions and perceptions
- Cognition works as a nested hierarchy of screens
- The innermost screen is phenomenologically transparent (Metzinger 2003)
- It's models all the way down to the inner layer of the nested hierarchy
- Return of the Cartesian theatre?

From: Ramstead & Albarracin 2023, preprint

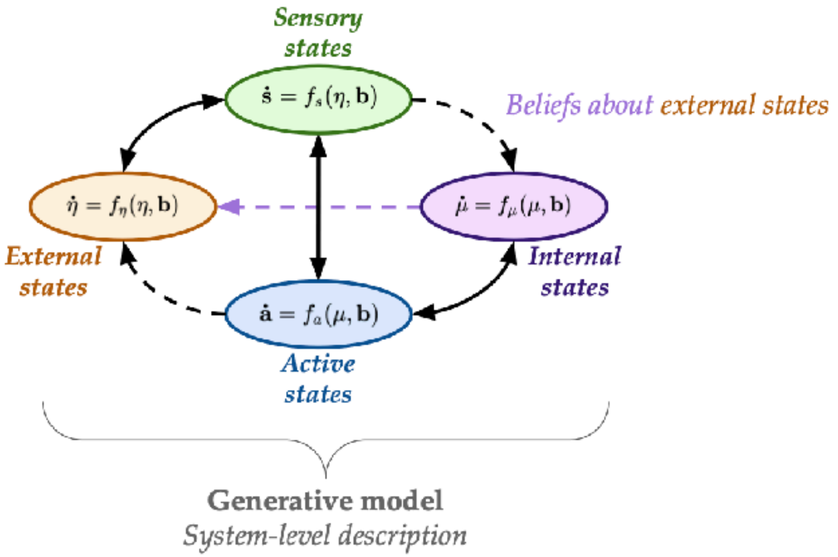
Free Energy as MUM?

- Free energy principle is empirically minimal, conceptually minimal but is it minimally unifying?
- It is so general it has different interpretations which come with different ontological commitments
- Criticism focuses on unclear empirical implementation and neurobiological mechanisms (Marvan & Havlik 2021)
- The free energy principle only provides a model structure for a plethora of models (Andrews 2021)

$$
\begin{aligned}
F[Q,y] &= \underbrace{-\mathbb{E}_{Q(x)}[\ln P(y,x)]}_{\textit{Energy}} - \underbrace{H[Q(x)]}_{\textit{Entropy}} \\
&= \underbrace{D_{KL}[Q(x) \,||\, P(x)]}_{\textit{Complexity}} - \underbrace{\mathbb{E}_{Q(x)}[\ln P(y|x)]}_{\textit{Accuracy}} \\
&= \underbrace{D_{KL}[Q(x) \,||\, P(x|y)]}_{\textit{Divergence}} - \underbrace{\ln P(y)}_{\textit{Evidence}}
\end{aligned}
$$

Free Energy as Model Template

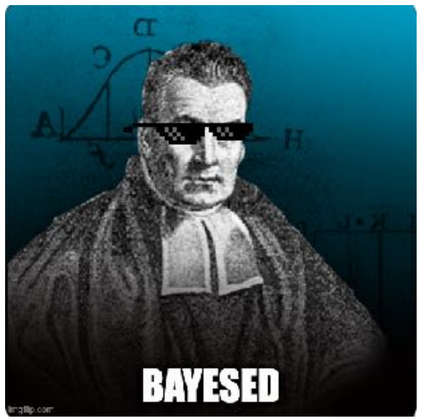
- Free energy has been used in physics, chemistry, AI, and cognitive science
- The free energy principle is unfalsifiable and needs to be interpreted and applied to cognitive systems
- Active inference is a process theory of how cognitive systems minimise free energy through perception and action
- High-level of abstraction and idealisation is a central feature of useful computational models
- Free energy principle unifies models not phenomena!



From: Ramstead & Albarracin 2023, preprint

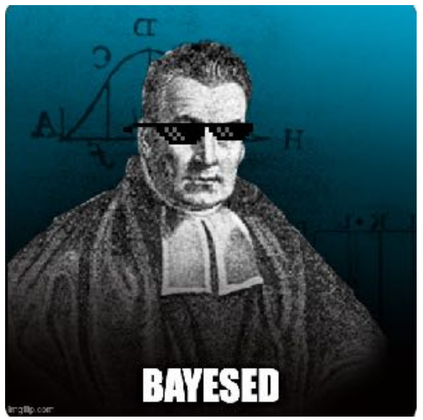
Conclusion

- Instead of coming up with more models of consciousness we need a common ground
- A minimally unifying model could help to translate ideas between different approaches
- Integration of First & Third Person approaches necessary
- Free energy models do not provide a MUM but a consensus on its interpretation could



Outlook

- Finding model templates in consciousness science could allow for better comparison of models
- The goal is to track overall structure of consciousness model landscape to build bridges and judge experiments
- It’s an interdisciplinary effort so we need to keep discussing and comparing our different approaches!



References

Knuuttila, Tarja, and Andrea Loettgers. 2023. “Model Templates: Transdisciplinary Application and Entanglement.” *Synthese* 201 (6): 200. https://doi.org/10.1007/s11229-023-04178-3.

Ramstead, Maxwell J. D., Anil K. Seth, Casper Hesp, Lars Sandved-Smith, Jonas Mago, Michael Lifshitz, Giuseppe Pagnoni, et al. 2022. “From Generative Models to Generative Passages: A Computational Approach to (Neuro) Phenomenology.” *Review of Philosophy and Psychology*, March. https://doi.org/10.1007/s13164-021-00604-y.

Marvan, Tomáš, and Marek Havlík. 2021. “Is Predictive Processing a Theory of Perceptual Consciousness?” *New Ideas in Psychology* 61 (April): 100837. https://doi.org/10.1016/j.newideapsych.2020.100837.

Seth, Anil K., and Tim Bayne. 2022. “Theories of Consciousness.” *Nature Reviews Neuroscience* 23 (7): 439–52. https://doi.org/10.1038/s41583-022-00587-4.

Humphreys, Paul. 2020. “Emergence: A Philosophical Account,” 30.

Baars, B. J. A Cognitive Theory of Consciousness (Cambridge Univ. Press, 1988).

Hameroff, Stuart, and Roger Penrose. 2014. “Consciousness in the Universe: A Review of the ‘Orch OR’Theory.” *Physics of Life Reviews* 11 (1): 39–78.

Weingarten, Carol P., P. Murali Doraiswamy, and Matthew P. A. Fisher. 2016. “A New Spin on Neural Processing: Quantum Cognition.” *Frontiers in Human Neuroscience* 10 (October). https://doi.org/10.3389/fnhum.2016.00541.

Kelly, Yin, Taylor Webb, Jeffrey Meier, Michael Arcaro, and Michael SA Graziano. 2014. “Attributing Awareness to Oneself and to Others.” *Proceedings of the National Academy of Sciences U. S. A.* 111 (13): 5012–17. https://doi.org/10.1073/pnas.1401201111.

Wiese, Wanja. 2020. “The Science of Consciousness Does Not Need Another Theory, It Needs a Minimal Unifying Model.” *Neuroscience of Consciousness* 2020 (1). https://doi.org/10.1093/nc/niaa013.

https://www.quantamagazine.org/what-a-contest-of-consciousness-theories-really-proved-20230824/

possiblelife.eu

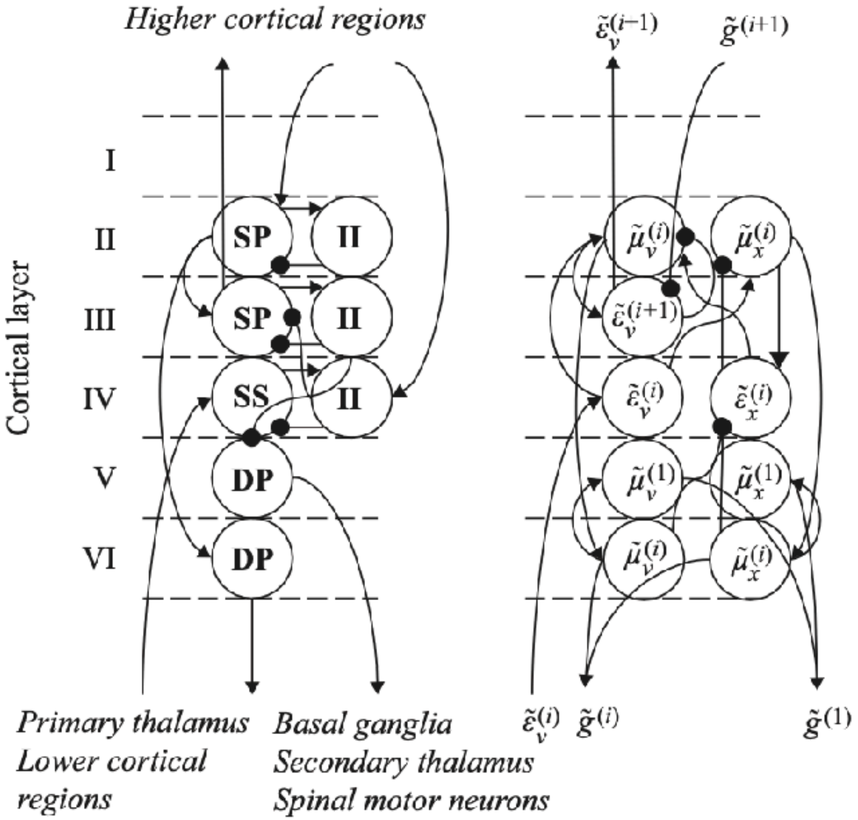
Abstract

- The scientific study of consciousness has always been an interdisciplinary effort combining theories and tools from many fields. Minimal models provide a shared framework for the systematic study of phenomena identifying common assumptions. With a plethora of consciousness models available, a critical analysis of overlaps and tensions becomes necessary to map out different approaches. Model templates provide a philosophical and computational tool to track how ideas travel between different theories and this approach could supplement the search for a minimal model of consciousness. In my talk, I present the model template approach and discuss how it can be used to understand the migration from physics to biology, psychology, and sociology.

- As a case study, I analyse the free energy principle and how it uses ideas from physics and information theory to explain consciousness. The free energy principle suggests that organisms strive to minimise their surprise or uncertainty about their internal and external states, which can be seen as a foundational principle of self-organisation and adaptive behavior. I defend a pragmatic philosophy when it comes to our use of mathematical tools to model consciousness and I discuss how free energy models fit in with current paradigms in the cognitive sciences.

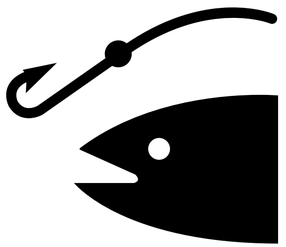
Neurobiology

- As a fundamental principle the free energy principle is not testable directly
- Only mechanisms derived from it can be applied to explain neurobiological systems
- Free energy models have been constructed to match the hierarchy of cerebral cortex
- Neurotransmitters encode parts of the model: dopaminergic systems is associated to precision of policies, serotonin precision of prior preferences (Parr & Friston 2018)



Model Templates

- Originates from ideas about theoretical and computational templates (Humphreys 2002)
- Tracks interdisciplinary use and transfer of mathematical models
- Model templates provide transdisciplinary common ground (Knuuttila & Loettgers 2023)
- What is transferred is not only the mathematical structure but also theoretical concepts
- Example: Lotka-Volterra model(s)

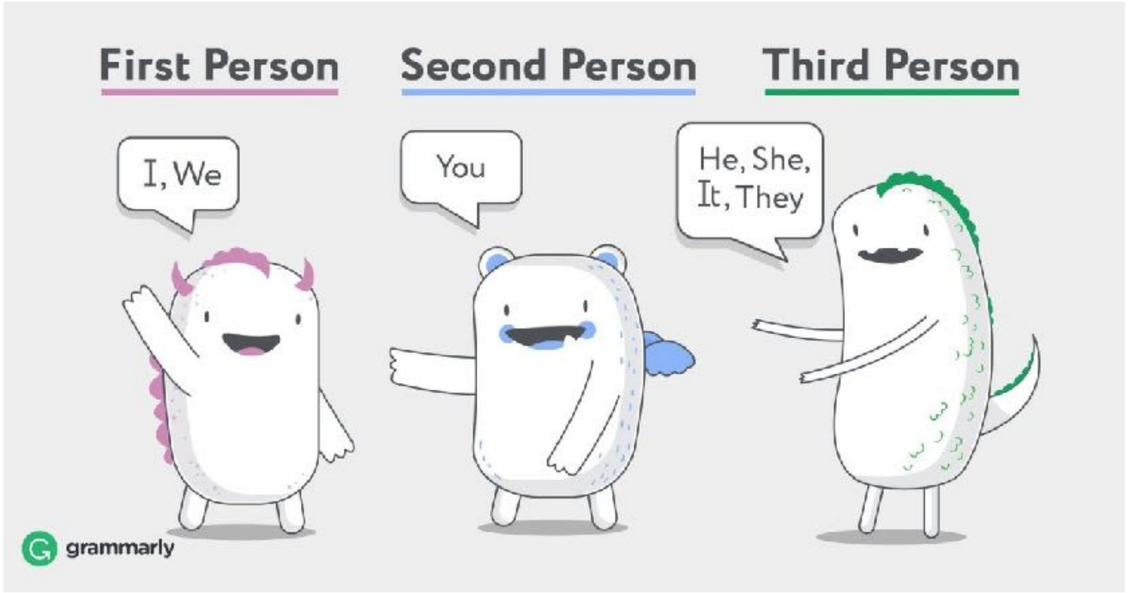


First Person Perspective

1st Person: Subjective experience

3rd Person: Scientific statements

"… giving an explicit and central role to first-person accounts and to the irreducible nature of experience, while at the same time refusing either a dualistic concession or a pessimistic surrender to the question"
Varela 1996, p. 333



From: grammarly.com

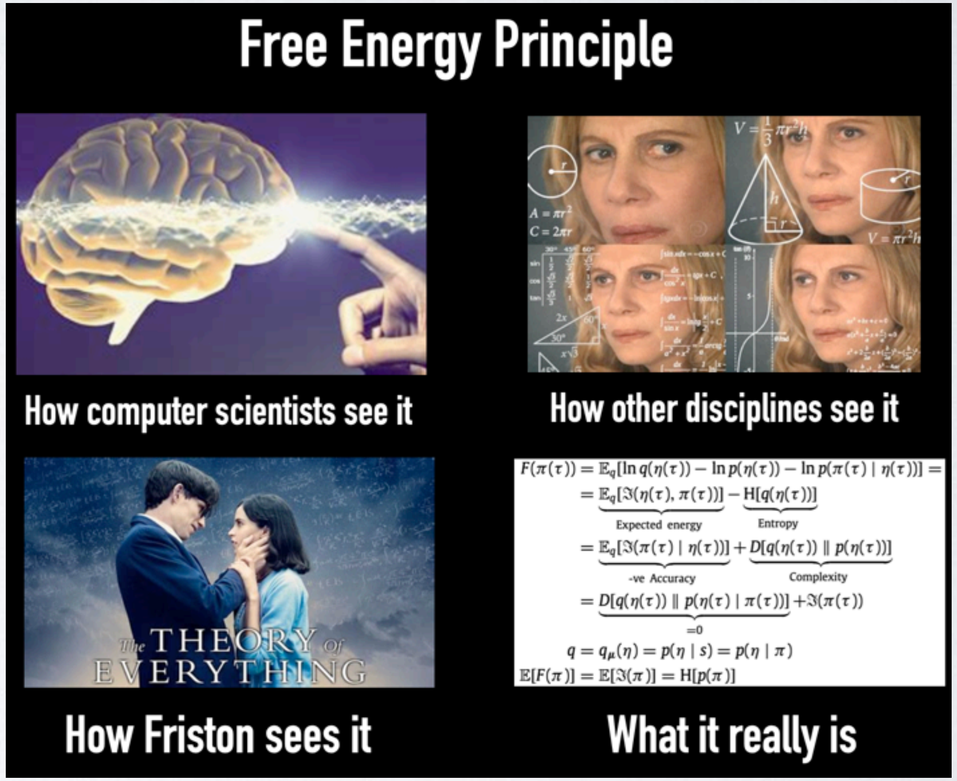
MARKOVIAN MONISM I

- What the free energy principle is in the literature:

  a grand unifying theory
  a modelling framework
  a modelling heuristic
  a new branch of physics
  a formal ontology

- What the free energy principle really is:

  a model structure without empirical content (Andrews 2021) or a model template?



MARKOVIAN MONISM 2

- So what are the philosophical foundations of the free energy principle?
- Markovian Monism:

  1. there is only one type of thing and only one type of irreducible property
  2. if systems have mental properties, then they have them partly by possessing a Markov blanket

- Accepts mechanism, physicalism, and modest representationalism (Friston et al. 2020)
- Rejects dualism, reductionism, and epiphenomenalism (Ramstead 2023)

THE STRANGE INVERSION

- A strange inversion of reasoning (Dennett 2009)
- Organisms must not minimise free energy to exist, but if they exist they can be modelled as minimising free energy
- Emergence through constraints: the whole is less than the sum of its parts

"the resulting philosophical perspective is not physicalist reductionism (a reduction of causal efficacy to "mere" physics)—but rather, a deep commitment to anti-reductionism"
Ramstead 2023

NON-REPRESENTATIONAL PERSPECTIVE

- Internal/External distinction only makes sense in relation to the boundary
- The process of separation is more important than the distinction itself
- Blanket dynamics are a dynamic coupling instead of parameters representing external states
- Going back to Varela and formalising phenomenology (Roy et al. 1999)

"because of the symmetric setup of the Markov blanket, it would be possible to repeat everything above but switch the labels of internal and external states—and active and sensory states—and tell the same story about external states tracking internal states."
Friston et al. 2023, p.11