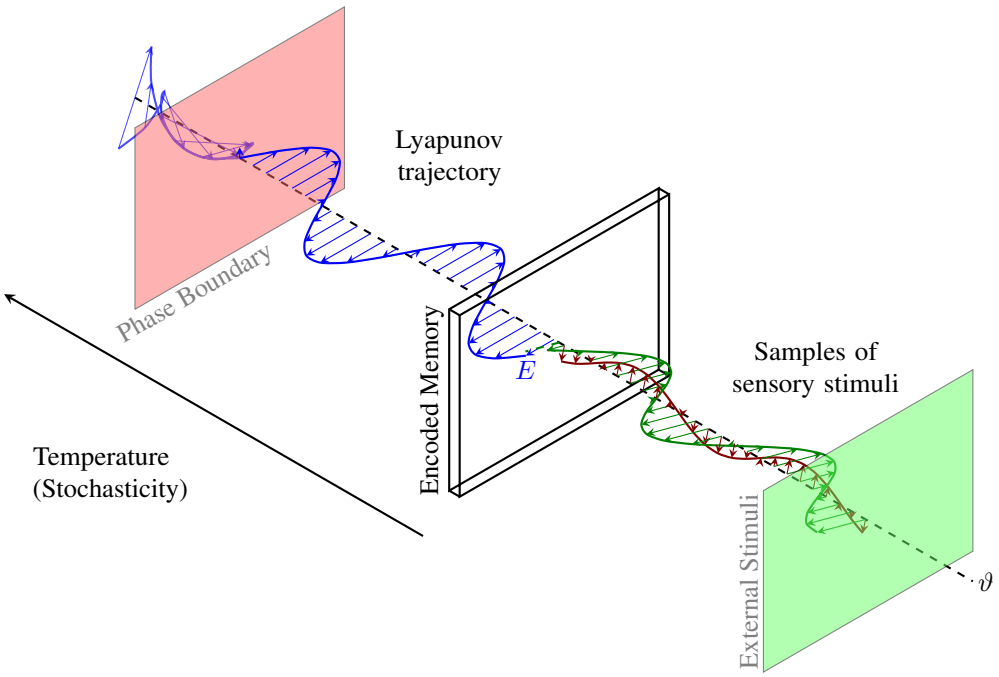
Energy-Based Dynamics in Consciousness

Luke J. Pereira

Abstract

The *free energy principle* shows us that biological dissapative systems act upon the environment to resist phase transitions. By learning to endure, an adaptive system avoids phase transitions that would otherwise change their physical structure. It is possible to define an adaptive artificial agent that reproduces this behaviour in a latent phase space that is bounded by two energy manifolds. The artificial agent can then learn how to maintain a dynamic trajectory in a phase space that minimizes the free energy of the joint energy functions represented as a Lyapunov function. Equivalently, this can be expressed as minimizing the Kullback–Leibler divergence or relative entropy of the joint probability densities. A comparison can be made between the agent's dynamic trajectory in latent space and the dreams, which can be understood as sensorimotor hallucinatory experiences that follow a narrative structure.



A depiction of the high-dimensional energy manifolds that act as phase boundaries between the states of consciousness.

1 Introduction

Energy-based models (EBMs) provide an alternative perspective to the standard optimization approach of using cost function to measure a model's ability to learn a probability density by providing a useful layer of abstraction to build on. I propose a phase space that is divided by two such manifolds described by separate energy functions that delimit phase boundaries of the three states in which the agent can exist. The approach of maintaining two energy functions is in contrast to the standard approach of minimizing a single energy function (the cost function). Having a distance between an accurate world mode and an upper bounded implausible reality allows for better predictions within a stochastic world model and creative experimentation and imaginative exploration in an agent's actions.

A dissipative system is a thermodynamically open system that exchanges energy and matter with an environment. There are several notions of a dissipative system, one being the existence of a Lyapunov function. By dynamically optimizing the internal parameters of an agent we minimize the free energy represented as a Lyapunov function or equivalently as the Kullback–Leibler divergence or relative entropy. At higher states, the temperature or stochasticity of the environment grows which decreases the stability of the dissapative agent and increases the likelihood that the agent crosses a phase boundary and transitions its state.

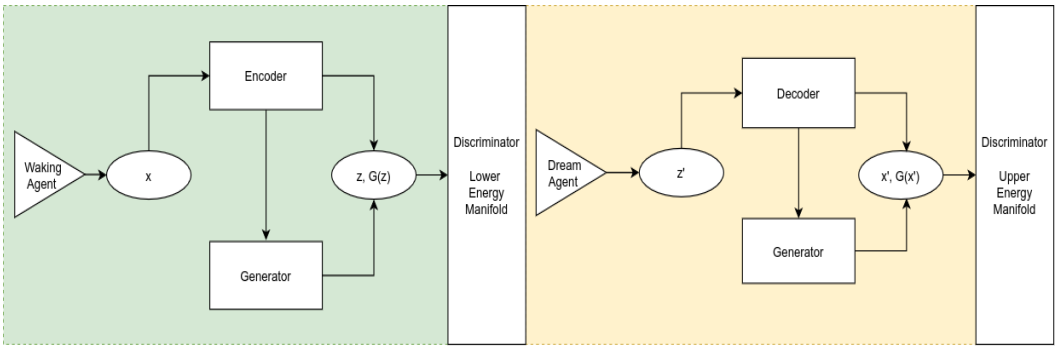
If we label the three states an agent can exist in as *waking* consciousness, *dreaming* consciousness, and *lucid dreaming* consciousness we establish a useful model of the mind. The lowest phase space is observable reality and the phase space between the two manifolds is the dream world. It is sufficiently distanced so that confabulation or prediction error does not push the agent into a phase transition. The lower phase boundary, dividing waking and dreaming consciousness, is the first encoded discriminator network's energy function, E_{world}. This can be interpreted as a collection of observations of reality from the training data and any realistic generated predictions. This manifold serves as the lower bound or ground truth of the agent's world model. The higher phase boundary, dividing dreaming and lucid dreaming consciousness, is the energy manifold of a highly stochastic generative model, E_{dream}. The degree of stochasticity corresponds to the information loss that results from encoding observations.

When we dream, we enter the phase space between the accurate world model and a world model of minimal plausibility. Here, the parameters of our internal states are in continuous flux in order to resist phase transitions into waking or lucid dreaming states. Lucid dreaming consciousness can be intepreted as a form of secondary consciousness and is a form of knowledge distillation of the ensemble density of the dream world and the memory of experienced reality. In such a state, an agent is aware of both the dream world plausibility and the constraints of reality. From this vantage point, a system can temporarily distill knowledge and explore abstractions of a highly stochastic reality.

By showing that this metaphysical structure exists and interpreting it as a model of our consciousness, we simultaneously open two paths of exploration in both the realms of extending human consciousness through dream research and as an alternative approach to developing human-like creativity in artificial agents.

2 Training the Energy Functions

In a standard training algorithm, we aim to minimize a loss from a cost function. In terms of our energy model, this is equivalent to minimizing the space between a generative world model to make it as close to the training data as possible. This results in an artificial intelligence agent that exists as closely in the waking state as possible but lacks an imaginative aspect that is crucial to for a creative consciousness the produces novelty. Instead, it is possible to simultaneously train an artificial agent to have both a realistic model of the world but be able to have an understanding of how to creatively modify its world model to some extreme. Moreover, the agent should be able to validate its modifications, store its ideas in memory and later decode and enact the feasible ideas on the external world environment.



There are two phases of training, which can occur in parallel or sequentially. In the waking phase, an agent explores the test environment and creates an accurate model of the latent world. It does this by first encoding its true sensory data, then producing synthetic predictions using the latent representation and training a discriminator on the pair of outputs. The discriminator will be the lower energy manifold that acts as a phase boundary between the waking and dreaming state. It is possible encoding is done with a Variational Autoencoder (VAE).

In the dreaming phase, an agent explores a training environment in latent space and creates and upper bound on sensory possibilities. It does this by first decoding its latent state, then generating a sensory interpretation of the state and training a discriminator on the pair of outputs. The dream model is able to produce highly implausible and incoherent data up to some stochastic threshold. This threshold can be proportional to the dimensional difference between the information bottleneck of the encoder and the original dimension of the input. The energy function of this secondary discriminator will serve as the upper manifold that represents the limits of imagination, beyond which the environment becomes incomprehensibly random and the agent becomes lucid.

In a sense, this approach is similar to the World Models architecture (Ha and Schmidhuber, 2020). However, instead of using a Mixture Density Network combined with a RNN (MDN-RNN) to predict a range of stochastic states, the trajectories of a dynamical system will be computed by minimizing free energy of a composition of these bounding densities in order to train the actions and sampling of the agent.

3 The Free Energy Principle

Biological systems are thermodynamically open, in the sense that they exchange energy and entropy with their environment. Furthermore, they operate far-from-equilibrium and are dissipative, showing self-organizing behaviour. However, biological systems are more than simply dissipative self-organising systems. They can negotiate a changing or non-stationary environment in a way that allows them to endure over substantial periods of time. This endurance means that they avoid phase transitions that would otherwise change their physical structure (Friston, 2006).

Let ϑ parameterize environmental forces or fields that act upon the agent and λ be quantities or temperature that describe the agents physical state. The free energy is a scalar function of the ensemble density and the current sensory input. Let $q(\vartheta;\lambda)$ be an arbitrary density function on the environments parameters that is specified or encoded by the agents parameters. It can be regarded as the probability density that a specific environmental state ϑ would be selected from an infinite ensemble of environments given the agents state λ, which is fixed and known. Then the free energy of the agent is given by,

$$\begin{aligned} F &= \int q(\vartheta) \ln \frac{p(\tilde{y},\vartheta)}{q(\vartheta)} d\vartheta \\ &= -\langle \ln p(\tilde{y},\vartheta) \rangle_q + \langle \ln q(\vartheta) \rangle_q \end{aligned} \tag{1}$$

Here $\langle \cdot \rangle_q$ means the expectation under the ensemble density q.

4 Training and Testing the Agent

A Lyapunov function is constructed to represent the composition of the two energy functions. A Lyapunov function is a scalar function of a systems state that decreases with time. Instead of trying to infer the Lyapunov function given an agent's structure and behaviour, we train the agent to minimize its Lyapunov function (its free energy) by optimizing its parameters. The free energy principle states that all the quantities that that are owned by the system will change to minimize free energy. These quantities are the agent's internal parameters λ and the action parameters α. We can rearrange (1) to show the dependence of the free energy on α and λ,

$$\begin{aligned} F &= -\ln p(\tilde{y}) + D(q(\vartheta;\lambda)||p(\vartheta|\tilde{y})) \\ &= -\langle \ln p(\tilde{y},\vartheta)\rangle_q + D(q(\vartheta;\lambda)||p(\vartheta|\tilde{y})) \end{aligned} \tag{2}$$

Where D is the Kullback–Leibler cross-entropy or divergence term that measures the difference between the ensemble density and the conditional density of the causes. Changing the configuration of the system to move or resample the environment by optimizing its actions α will minimize the free energy of the first term. The Kullback–Leibler is used to descend the Lyapunov free energy by optimizing the agents internal parameters λ in the second term with,

$$D(q(\vartheta;\lambda) = \int q \ln \frac{q}{p} d\vartheta.$$

During the waking phasing, the agent now has access to both the discriminators and generators of the world model and the upper bounded dream model. It can simultaneously use them to predict and imagine in order to make the best decisions over a wider range of possible states.

5 Validation Experiments

It is conceivable that a dissapative state of lucid dreaming in humans can be extended using an external system to provide perturbations that drive a dreaming consciousness to transition phases into lucid dreaming consciousness and stabilize in this state. In this state we would be able to freely modify our world without being constrained by physical reality or our physical abilities. From here, it may be possible to store and later decode our latent actions and ideas in reality.

It is also conceivable that the state of a dreaming artificial intelligence agent can be pushed to transition phases into lucidity, which is a higher analog of the shallow consciousness experienced when being trained solely on sensory inputs. It's unknown what role dream lucidity plays in the development of consciousness or creativity. In a simulation game, we may attempt to use lucidity of a computer to unveil tactics or strategies that we were unaware of in order to improve our performance and our understanding of the game. It can be posited that the game of consciousness is the interplay between man and machine within nested simulation.

References

[1] Friston, K.J., Stephan, K.E. Free-energy and the brain. Synthese 159, 417–458 (2007).

[2] D. Ha and J. Schmidhuber. World models. In Advances in Neural Information Processing Systems (NIPS), 2018.

[3] Friston KJ, Daunizeau J, Kiebel SJ (2009) Reinforcement Learning or Active Inference?. PLoS ONE 4(7): e6421. doi:10.1371/journal.pone.0006421

[4] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[5] Nir Y, Tononi G. Dreaming and the brain: from phenomenology to neurophysiology. Trends Cogn Sci. 2010;14(2):88-100. doi:10.1016/j.tics.2009.12.001