Energy-Based Dynamics in Consciousness

1 Introduction

The *free energy principle* shows us that biological dissapative systems act upon the environment to preclude or resist phase transitions. This endurance means that an adaptive system avoids phase transitions that would otherwise change its physical structure. It is possible to define an artificial adaptive agent that can replicate this behaviour in a metaphysical phase space. I propose a phase space that is divided by two manifolds described by separate energy functions that delimit phase boundaries of the three states in which the agent can exist. 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.

We can label the three states an agent can exist in as *waking* consciousness, *dreaming* consciousness, and *lucid dreaming* consciousness. 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 transitions its state into a lower state. The approach of maintaining two energy functions is in contrast to the standard approach of minimizing a single energy function (i.e. a cost function) that represents the density of the training data. Inference is instead performed by descending the combination (i.e. a mixture, product, or composition) of the densities.

The lowest phase space is observable reality and the phase space between the two manifolds is the dream world. It is sufficiently distanced, that is, the energy from one to the other is large enough so that confabulation or prediction error does not push the agent into a phase transition. Having a distance between observed reality and a minimal plausible reality allows for predictions of a stochastic world model and creative experimentation in the agent's actions.

The lower phase boundary, dividing waking and dreaming consciousness, is the first encoded discriminative network's energy function, E_{lower}. This can be interpreted as a collection of observations of reality (i.e. the training data) and the realistic generated predictions that have created impressions in the manifold. E_{lower} 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_{upper}.

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, our actions, and the environment are in continuous flux in order to resist phase transitions into waking or lucid dreaming states by minimizing the joint free energy of the two energy functions. There are several notions of a dissipative system, one being the existence of a Lyapunov function. This dynamic optimization of the parameters of consciousness correspond to minimizing free energy in terms of the Kullback–Leibler divergence or relative entropy.

I propose that lucid dreaming consciousness is a form of secondary or higher 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, we simultaneously open up two paths of exploration in both the realms of extending human consciousness through dream research and as an alternative approach to developing human-like consciousness in artificial intelligence research.

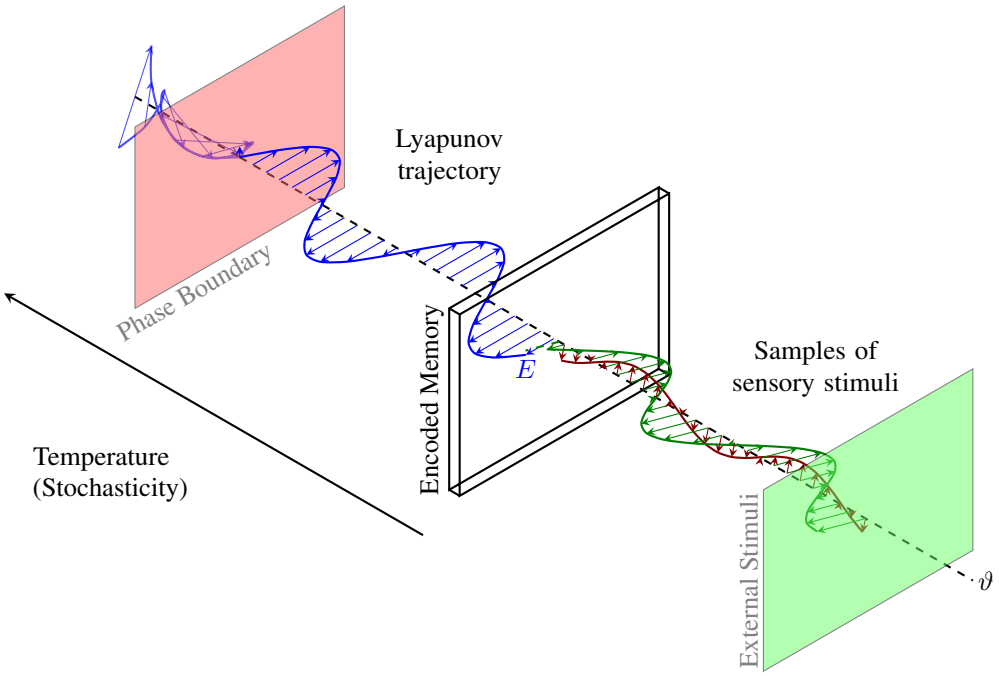
2 The Free Energy Principle

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.

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.



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

3 Training the Energy Functions

A Bidirectional GAN (BiGAN) composed of an autoencoder, a generator, and a discriminator is trained in parallel to learn an accurate depiction of the world from both the training and synthetic data. This will be the lower energy manifold that acts as a phase boundary between the waking and dreaming state. It also is possible encoding is done with a Variational Autoencoder (VAE).

A secondary discriminator model is trained from the encoded real and synthetic data but 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. A GAN or Variational Autoencoders (VAE) can be used for this.

In a sense, this approach is similar to the World models architecture. However, instead of using a Mixture Density Network combined with a RNN (MDN-RNN), the trajectories of a dynamical system are computed by minimizing free energy of a composition of densities in order to train the actions and environment sampling taken by the agent.

4 Training 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,

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

Trajectories from throughout the space can be simulated in parallel and equilibrium points can be analyzed.

5 Knowledge Distillation of Multi-Agents

The space of dreams may be vast and largely not very useful. Hinton makes an interesting analogy between an ensemble of multiple neural networks being trained on specific features with the composition of matter in the universe. He shows that a large portion of the information being stored in the ensemble of networks is actually redundant and is a kind of dark knowledge. In order to minimize the amount of computation and the model's memory footprint, this dark knowledge can distilled by training a new model using a softmax or logits of all the models. These operations reduce the range of information stored in the outputs while maintaining accuracy of relative dependencies in the distilled model.

It is conceivable that a similar process occurs in our minds during dreaming sleep where the single agent we understand as our base consciousness replicates its core parameters into a multi-agent representation. Instead of casting out single fishing rods to search the phase space, we can cast out an entire net. This allows us to simulate multiple latent trajectories and better explore the phase space in search of equilibrium points of interest. Since latent computations are less expensive, it is conceivable for the mind to be able to run multiple latent trajectories in parallel. That is, our base

consciousness periodically switches between its replicated agents by selecting agents with Lyapunov trajectories that have minimal positive energies with $E > 0$ and are strictly decreasing with $\frac{dE}{dt} < 0$. This enables the selection of agents that are nearest to approaching an equilibrium point using a variation of initial conditions in the environment. These points are of interest because they capture highly stochastic yet highly plausible states of the world model that haven't yet been experienced.

6 Validation Experiments

In a standard training algorithm, we aim to minimize loss using a cost function. In terms of our energy model, this is equivalent to compressing the space between the manifold so that our generative world model is as close to our 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 both have a realistic model of the world but be able to have an understanding of how to creatively modify its world model. Moreover, the agent should be able to validate its modifications, store its ideas in memory and later decode and enact the idea on the external world environment.

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 stabalize 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.

7 Appendix

7.1 Energy-Based Models

See EBM section of the *{Artificial, Biological} Neural Networks* notebook for more information:

`https://lukepereira.github.io/notebooks/documents/2020-neural-nets/main.pdf`

7.2 Dream Research

Dreams can be described as sensorimotor hallucinatory experiences that follow a narrative structure. There is some evidence that dreams, like output from generators, are also created in a top-down manner. That is, they originate in abstract knowledge and figurative thought and are then processed back into imaginal copies of perceptual phenomena. Dream recall was found to correlate best with abilities of mental imagery rather than language proficiency.

Children under the age of 7 reported dreaming only 20% of the time when awakened from REM sleep, compared with 80–90% in adults. Visuo-spatial skills are known to depend on the parietal lobes, which are not fully myelinated until age 7. Lesion studies also find a cessation of dreaming follows from damage in or near the temporo-parieto-occipital junction. Some lesions are associated with increased frequency and vividness of dreams and their intrusion into waking life, especially those in medial prefrontal cortex, the anterior cingulate cortex, and the basal forebrain.

In the Block Design Test of the Wechsler intelligence test battery, children look at models or pictures of red and white patterns, and then recreate those patterns with blocks. Scores on this test are the one parameter that correlates best with dream report in children. Children with the most developed mental imagery and visuo-spatial skills (rather than verbal or memory capabilities) report the most dreams.

In the mental rotation test, a subject is asked to determine whether two figures are the same or different given images of their rotated states. To perform well in this test, the subject needs to have a well-developed and functioning visuo-spatial and mental imagery skills since the act of rotating the object is a form of imagining.

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