# Intention is Physical

Natesh Ganesh
Dept. of Electrical and Computer Engineering
Univ. of Massachusetts, Amherst
Email: nganesh@umass.edu

## I. Introduction

How do we go from impersonal physical laws to intentions and goals? From an engineering perspective, this question is of great relevance now, given the growing interest in achieving artificial general intelligence. It will be very hard to build agents capable of matching or even exceeding human intelligence, if we cannot answer fundamental questions about agency. This essay will address some of these questions. I will begin by presenting the minimal dissipation hypothesis, followed by a short description of the methodology and the exact relationship between dissipation and learning dynamics. I will then extend these ideas to explain how goal-oriented agency would emerge in physical systems under the hypothesis, and present results from non-equilibrium thermodynamics to suggest a way to unify learning with the biological evolutionary process. I will conclude the essay with some exciting conjectures on connections to the critical brain hypothesis, and cognition as phase transitions in input mapping. I would like to make clear that while mathematical descriptions of the laws of nature might be a product of human cognition, I will work under the premise that the evolution of physical systems under these laws is ontologically objective. The familiar language of mathematics is a necessary evil to explain how subjective goals and intentions arise in physical systems.

## II. A Physical Basis for Learning: The Minimal Dissipation Hypothesis

I have been studying the fundamental limits on energy efficiency in new computing paradigms using physical information theory as part of my dissertation. Given the increase in machine learning applications, determining efficiency limits for physical instantiations of neural networks are of particular interest to me. The large number of algorithms, often arbitrary cost functions and non-intuitive training processes left me perplexed and looking for a clear basis for learning in physical systems from fundamental laws. In my search for a solution, I posed the following statement that I will refer to as the **minimal dissipation hypothesis** -

*Open physical systems with constraints on their finite complexity, that dissipate minimally when driven by external fields, will necessarily exhibit learning and inference dynamics.*

Are physical systems energy efficient because they learn, or are systems that dissipate minimally (and are highly energy efficient) in their environment, necessarily show dynamics that we identify as learning? This might just seem like a clever play of words, but let us take a second and think about it. Our numerous attempts at machine learning, often requiring tremendous resources have shown us that it is clearly possible to learn without being energy efficient. Can minimal dissipation alone be a sufficient condition for learning? Before I proceed, note that energy efficiency as an underlying principle in neural dynamics has been suggested before, but we will make clear the connection here [1].

In order to test this idea, I generalized the fundamental bounds on dissipation developed for finite state automata (FSA). The bounds on dissipation that are used in this paper are rigorously

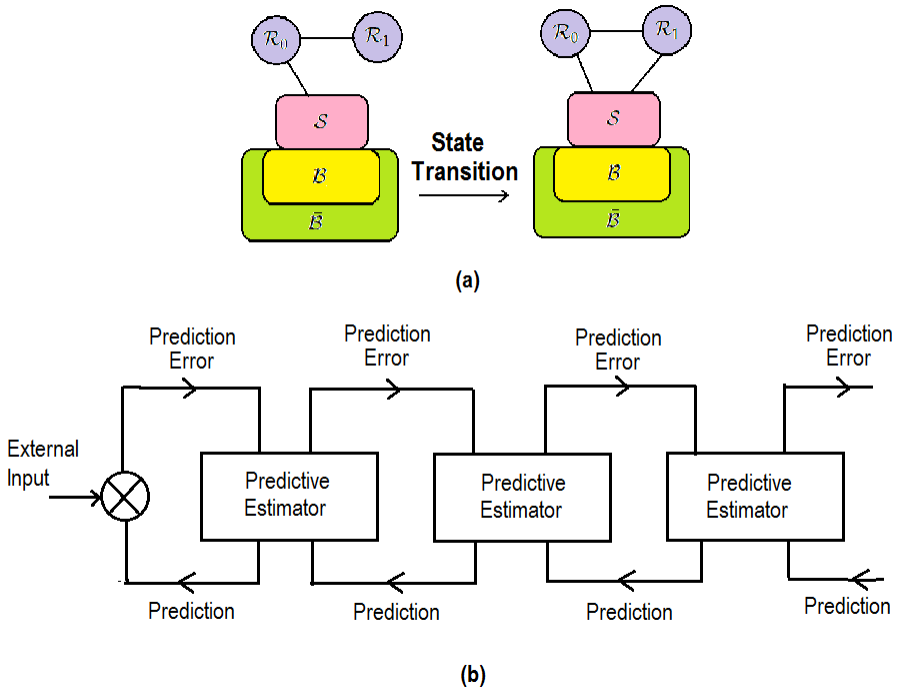


Figure 1: (a) The physical FSA instantiated in the states of system $\mathcal{S}$ that is correlated to the input string $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1$ and in contact with heat bath $\mathcal{B}$, undergoing a state transition. (b) Diagram of an hierarchical predictive coding architecture. Predictions travel from higher to lower levels, while prediction errors travel the other way around.

obtained from the dynamics of the system, and details on the corresponding methodology are available in [3]. A detailed physical description of an abstract and physical FSA can be found in [4], but I will present a quick summary here. The physical implementation of an abstract FSA is given as $\mathcal{F}_\mathcal{P} = \{\mathcal{S}, \mathcal{R}, \{\sigma^\mathcal{S}\}, \{\hat{x}^\mathcal{R}\}, \{\tilde{\mathcal{L}}\}\}$. The FSA is faithfully instantiatied in the orthogonal distinguishable states $\{\hat{\sigma}^\mathcal{S}\}$ of a quantum system $\mathcal{S}$. The external input strings $\{\hat{x}^\mathcal{R}\}$ that produce transitions in the FSA are instantiated in the physical 'referent' system $\mathcal{R} = \mathcal{R}_0\mathcal{R}_1$, with $\mathcal{R}_0$ corresponding the current state of the FSA and $\mathcal{R}_1$ for the latest, that is about to produce the next state change. We allow $\mathcal{R}_0$ and $\mathcal{R}_1$ to be correlated. The state transition (defined by transition mapping $\{\tilde{\mathcal{L}}\}$) is realized by the dynamical evolution of the state of $\mathcal{S}$, conditioned on the state of $\mathcal{R}_1$ (they exhibit Markov property) and in interaction with $\mathcal{B}$, a heat bath at temperature $T$ that is in contact with $\mathcal{S}$ (as shown in Fig.(1a)). Global evolution of the interacting composite $\mathcal{R}_1\mathcal{S}\mathcal{B}$ producing this transition is assumed to be governed by the time-dependent Schrodinger equation to ensure consistency with physical law (implying unitary evolution of the state of $\mathcal{R}_1\mathcal{S}\mathcal{B}$). The input $\mathcal{R}_1$ remains unchanged at the conclusion of the FSA state transition.

For the FSA $\mathcal{F}_\mathcal{P}$, the input-averaged amount of energy dissipated to a thermal environment $\Delta\langle E^\mathcal{B}\rangle$, on each state transition is lower bounded as

$$\Delta\langle E^\mathcal{B}\rangle \geq k_B T \ln 2 \left[-\Delta S^\mathcal{S} + \mathcal{I}^{\mathcal{R}_1\mathcal{S}'} - \mathcal{I}^{\mathcal{R}_1\mathcal{S}}\right]$$

where $k_B$ is the Boltzmann constant, $T$ is the temperature of the bath and $-\Delta S^\mathcal{S}$ is the loss of von Neumann entropy of the system $\mathcal{S}$ undergoes the state transition. $\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$ and $\mathcal{I}^{\mathcal{R}_1\mathcal{S}'}$ is the quantum

mutual information between system $\mathcal{R}_1$ and $\mathcal{S}$ before and after the state transition [1]. We can view $\mathcal{I}^{\mathcal{R}_1\mathcal{S}}$ as a measure of "prediction" of $\mathcal{R}_1$ by $\mathcal{S}$. A similar bound has been derived under a different set of assumptions in [5]. Since a number of biochemical processes in living systems operate near the limits of dissipation, these bounds are good approximations of the actual dissipation for these processes [6], and analysis based on them are very relevant. Furthermore, the bounds are derived from fundamental physical law and applies to all systems, providing us with valuable insight that is independent of implementation details.

I will assume that the states $\{\hat{\sigma}^{\mathcal{S}}\}$ of $\mathcal{S}$ are pure states, allowing me to reduce the von Neumann entropy to it's classical Shannon equivalent, and replace the quantum mutual information with Shannon mutual information. Let $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ indicate the probability that the $i$-th string of $\mathcal{R}_0$ maps to the $k$-th current state of $\mathcal{S}$ at time $t$. We can find out the required optimal mapping for which dissipation is minimized by performing an optimization of $\Delta\langle E^{\mathcal{B}}\rangle$ subject to a constraint on the complexity $\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$ [7], with respect to $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$. We construct the necessary Lagrangian $\mathcal{L} = \Delta\langle E^{\mathcal{B}}\rangle + \beta\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$, where $\beta$ is the tradeoff parameter between dissipation and complexity

$$\beta = \frac{\partial\Delta\langle E^{\mathcal{B}}\rangle}{\partial\mathcal{I}^{\mathcal{R}_0\mathcal{S}}}$$

The optimization problem can be described as

$$min_{p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})} \;\; \mathcal{L} = \Delta\langle E^{\mathcal{B}}\rangle + \beta\mathcal{I}^{\mathcal{R}_0\mathcal{S}}$$

We would need to solve the equation $\frac{d\mathcal{L}}{dp(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})} = 0$. Under the assumption that the next state transition matrix is homogenous and independent of past encodings i.e $\frac{dp(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}},j^{\mathcal{R}_1})}{dp(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})} = 0$, we would get a system of equations for $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$, $p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})$ and $p(k_t^{\mathcal{S}})$, that are parametricized using $\beta$ and need to be solved consistently [8]. We have the equation for $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ to be

$$p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0}) \propto e^{\frac{-1}{\beta}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})]-\ln p(k_t^{\mathcal{S}})\}}$$

where $D_{KL}[a|b]$ is the Kullback-Liebler (KL) divergence between distributions $\{a\}$ and $\{b\}$ [2]. If $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})]$ is low, $p(k^t|i^{\mathcal{R}_0})$ is higher and $i^{\mathcal{R}_0}$ is likely to map to a state $k_t^{\mathcal{S}}$ that better predicts $j^{\mathcal{R}_1}$, forming the basis of predictive inference in the dynamics. For equally predictive encodings, $\ln p(k_t^{\mathcal{S}})$ ensures preference of sparse mappings over dense ones [9].

Since we require the FSA system $\mathcal{S}$ to continue to be in a minimally dissipative state, we can obtain the transition probabilities $p(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}}, j^{\mathcal{R}_1})$ from earlier results. We find that for systems that continue to be in minimally dissipative state, the favored state transitions are those from $k_t^{\mathcal{S}}$-th state at time $t$, to $l_{t+1}^{\mathcal{S}}$-th state at time $t+1$ by the $j$-th input of $\mathcal{R}_1$, if the system can predict the next incoming input $\mathcal{R}_2$. In addition to assuming that $\frac{dp(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}},j^{\mathcal{R}_1})}{dp(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})} = 0$, if we assume the system is in a steady state and $-\Delta S^{\mathcal{S}} = 0$, we are left with the results from the Information Bottleneck (IB) method [8]. Like the IB method, the above equations can be implemented in physical systems through an hierarchical predictive coding model, similar to the one constructed by Rao and Ballard for the visual cortex [10], and by Friston within the Free-Energy Principle [11] (Fig.(1b)).

However if we let $\frac{dp(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}},j^{\mathcal{R}_1})}{dp(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})} \neq 0$, we can follow the same procedure as before and obtain a revised expression for $p(l_{t+1}^{\mathcal{S}}|k_t^{\mathcal{S}}, j^{\mathcal{R}_1})$ as follows

[1] The von Neumann entropy is the quantum analogue of the classical Shannon entropy, and depends upon the probabilistic description of the states of the system. The quantum mutual information between two systems is the quantum analogue of Shannon mutual information, and a measure of the amount of correlation between two systems.

[2] The KL divergence $D_{KL}[a|b]$ between distributions $\{a\}$ and $\{b\}$ is a measure of the non-symmetric difference between the two probability distributions and is lower if the two distributions are similar. $D_{KL}[a|b] = 0$ when $\{a\} = \{b\}$.

$$p(l^{\mathcal{S}}_{t+1}|k^{\mathcal{S}}_t, j^{\mathcal{R}_1}) \propto e^{\frac{-1}{\beta_1}\{D_{KL}[p(m^{\mathcal{R}_2}|i^{\mathcal{R}_0}j^{\mathcal{R}_1})|p(m^{\mathcal{R}_2}|l^{\mathcal{S}}_{t+1})]\}} e^{\frac{1}{\beta_2}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t)]\}}$$

In addition to first term from the previous case, there is the $e^{\frac{1}{\beta_2}\{D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0})|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t)]\}}$ term, which implies that if the prediction of $\mathcal{R}_1$ by $k^{\mathcal{S}}_t$ was poor and corresponding KL divergence was large, it will favor mappings away from $k^{\mathcal{S}}_t$ to $l^{\mathcal{S}}_{t+1}$. This is using past encodings to improve predictions by avoiding recent errors i.e an error correction mechanism. These results show that learning dynamics are inevitable in a trade off between energy dissipation and statistical complexity, and give us a solid base to expand on and tackle issues regarding agency and goals.

## III. Intentional Agency

It is important to start with a clear and simple definition for agency. Agency is the capacity of a system/entity/agent/organism to act on it's environment. It can be broadly classified as unintentional (unconscious and involuntary), and purposeful goal-directed intentional agency. We have goals as desirable results of an action. It is the latter type of agency that we are particularly interested in, especially when the intentions and objectives are generated by the agent directly, rather than have it bestowed on them externally. We will define sense of agency (SA) as the pre-reflective subjective awareness that one is initiating, executing, and controlling one's own volitional actions in the world. In normal, non-pathological experience, the SA is tightly integrated with one's sense of ownership (SO), which is the pre-reflective awareness that one is the owner of an action.

In this section, I will explain the emergence of goal-directed agency in systems that dissipate minimally. We will now consider a joint quantum system $\mathcal{SA}$, which behaves like a FSA, but can also act upon it's environment and affect the next incoming input $\mathcal{R}_1$. System $\mathcal{SA}$ is defined as before, with actions dependent on the the states of $\mathcal{A}$ capable of affecting which input is seen next by the joint system (I am imbuing system $\mathcal{A}$ with agency, but not with a specific goal or purpose). We place no restriction on the nature of interaction between systems $\mathcal{S}$ and $\mathcal{A}$. Following the procedure in [4], we can calculate the lower bound on dissipation in this system as it undergoes a state transition -

$$\Delta\langle E^{\mathcal{B}}\rangle \geq k_B T \ln 2 \left[-\Delta S^{\mathcal{SA}} + \mathcal{I}^{\mathcal{R}_1\mathcal{SA}'} - \mathcal{I}^{\mathcal{R}_1\mathcal{SA}}\right]$$

We construct the required Lagrangian $\mathcal{L} = \Delta\langle E^{\mathcal{B}}\rangle + \beta\mathcal{I}^{\mathcal{R}_0\mathcal{SA}}$ and optimize it with respect to $p(k^{\mathcal{S}}_t l^{\mathcal{A}}_t|i^{\mathcal{R}_0})$. We see that in the optimal encoding, we have that

$$p(k^{\mathcal{S}}_t l^{\mathcal{A}}_t|i^{\mathcal{R}_0}) \propto e^{\frac{-1}{\beta}D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0},l^{\mathcal{A}}_t)|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t,l^{\mathcal{A}}_t)]} e^{\frac{1}{\beta}D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0},l^{\mathcal{A}}_t)|p(j^{\mathcal{R}_1})]}$$

From the above expression, we can see that as before we have that $p(k^{\mathcal{S}}_t l^{\mathcal{A}}_t|i^{\mathcal{R}_0})$ is higher when $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l^{\mathcal{A}}_t)|p(j^{\mathcal{R}_1}|k^{\mathcal{S}}_t, l^{\mathcal{A}}_t)]$ is lower, implying that $i^{\mathcal{R}_0}$ is likely to map to that $\mathcal{SA}$ state that makes good a prediction of the future distribution as predicted by that state of $\mathcal{A}$ (and corresponding action) and input $\mathcal{R}_0$. $p(k^{\mathcal{S}}_t l^{\mathcal{A}}_t|i^{\mathcal{R}_0})$ is also increased if $D_{KL}[p(j^{\mathcal{R}_1}|i^{\mathcal{R}_0}, l^{\mathcal{A}}_t)|p(j^{\mathcal{R}_1})]$ is higher, implying the future distribution produced by the state of $\mathcal{A}$ deviate from the mean distribution. The optimal encoding of $\mathcal{R}_0$ in $\mathcal{SA}$ is a trade-off between exploiting known information and exploration. Perhaps *wandering* is not such a bad thing.

While the state of $\mathcal{A}$ depends upon balancing exploration with prediction, the state of system $\mathcal{S}$ given the state of $\mathcal{A}$ needs to be maximally predictive of this future input. The state of the system $\mathcal{SA}$ corresponding to best future input prediction $p(j^{\mathcal{R}_1}_{t+1}|k^{\mathcal{SA}}_t)$ can form the basis of aims and intentions. Let us consider the following example case using the human brain, and say systems $\mathcal{S}$ and $\mathcal{A}$ correspond to the neocortex (responsible for sensory perception) and motorcortex

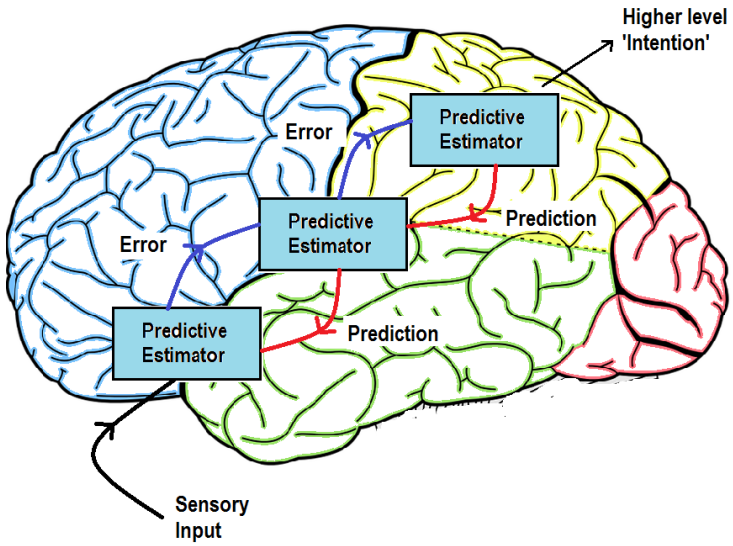


Figure 2: Diagram of higher level predictions in the hierarchical model corresponding to intentions.

(responsible for voluntary movement) respectively, and employ it in this example. Let the system $\mathcal{SA}$ be affected by the incoming visual input of say a basketball. The state of the joint system that predicts the future input depends on how the past inputs are encoded $p(k_t^{\mathcal{SA}}|i^{\mathcal{R}_0})$, lending weight to the adage that *our aims and goals are shaped by our history*. Due to these past inputs, let the state of the system $\mathcal{A}$ (motor cortex) that is most likely given the prediction-exploration trade off, corresponds to the action "throw the ball". Since the state of $\mathcal{S}$ (neocortex) has to be correlated with $\mathcal{A}$ in order to maximally predict the future input of "ball being thrown", the states of $\mathcal{S}$ corresponding to sensory perception make predictions for "seeing the ball being thrown". Thus the joint state of $\mathcal{SA}$=("see ball being thrown","throw ball") as the ball is thrown will explain the sense of agency, the awareness of an action being performed as it being performed.

We now move from the sense of agency to the actual intention of an action. We should note that not all levels in an hierarchical model is making predictions about the exact same features. For example, in the case of visual perception of a face, the higher levels make predictions corresponding to say, "seeing a face," which is then translated with greater details on features like mouth and eyes and nose, as it is passed into the lower levels (Fig.(2)). Similarly predictions made in the higher levels of the hierarchical model in $\mathcal{SA}$, under the minimal dissipation hypothesis, would correspond to the higher level intention of the action-sense of agency (like say "win game" in our example), which would then be translated with greater detail downstream to the lower level subsystems, that interact directly with the external input. If the prediction is correct i.e. confirming at the lower level with the incoming input of "ball has been thrown," and in the form of "game won" in the higher level, the system perceives as the objective being successfully accomplished. While I might have distinguished between features of perception vs intentional action, generated in the upper levels of the hierarchical model, arguments have been made for inherent intentionality in every perception event [12]. We can view the upper levels of the hierarchical model in the brain as the source of only intentions and make a strong case that *intention is physical*.

Crucial to this process though, is a sense of ownership that the system will learn over time about what is within the system's control and what is beyond that. If the system initially predicts a future input that is beyond what is controllable by it's own actions, then in time the minimal

dissipation hypothesis will ensure that system can learn to not make that erroneous prediction. While we have specified a larger principle of minimal dissipation to explain aims and goals in physical system, the exact state transition mappings will be heavily dependent on the physical system in question and can vary significantly. It is possible for two different physical systems, both that dissipate minimal to make completely different predictions of the future (manifesting as having different goals and aims), even if they have the same history of inputs. In fact it is possible for the same structure with a fixed history, to have different predictions at a certain time. This symmetry breaking at critical bifurcation points will be explained in detail later. Deeper meaning and hidden objectives to our actions can always be obtained by a reflective analysis after the fact, and are beyond the scope of this essay. While examples using the human brain are easy to understand, the results themselves do not assume the requirement of a brain of any sort to be present.

I used the minimum dissipation hypothesis to explain the emergence of goals and intentions, but started the section by assuming agency in the system (without imbuing it with a goal or purpose). The intention and objectives of the agent arose only from the objective physical laws depending upon the system under consideration. In the next section, I will present results from non-equilibrium thermodynamics to explain how agency might have arose in physical systems in the first place, and tie up the idea with the dissipation driven adaptation hypothesis.

## IV. An Argument from Non-Equilibrium Thermodynamics

The results from the previous sections do not depend on the following arguments but definitely did reaffirm the idea of systems that dissipate minimally. Jeremy England presented the dissipation driven adaptation hypothesis using a revised version of Crook's fluctuation [14].

$$\frac{\pi(i\to j)}{\pi(i\to k)} = \frac{\pi(j^*\to i^*)}{\pi(k^*\to i^*)} \frac{\langle e^{\frac{-\Delta Q}{k_B T}}\rangle_{i\to k}}{\langle e^{\frac{-\Delta Q}{k_B T}}\rangle_{i\to j}}$$

From the above equation, all things being equal, we are more likely to find a system in state $i$ than in $j$ ($\pi(i\to j) > \pi(i\to k)$), if $\langle e^{\frac{-\Delta Q}{k_B T}}\rangle_{i\to k} > \langle e^{\frac{-\Delta Q}{k_B T}}\rangle_{i\to j}$. The author argues for the dissipation driven adaptation hypothesis, in which while most changes in system configurations are random when driven by external fields, the most irreversible of these changes are ones that have better work absorption and dissipation structures for that driving field. In time, the history of these specific configurations add upto to produce a structure that appears adapted to it's environment. I will ascribe the emergence of agency (but not intentions) in physical systems as a result of such dissipation driven adaptation. The apparent conflict between this hypothesis and the minimal dissipation hypothesis can be resolved as follows.

England derived the fluctuation theorem for macrostate-to-macrostate transitions [13]

$$\frac{\pi(I\to II)}{\pi(II\to I)} = \left\langle \frac{\langle e^{-\beta\Delta Q_{i\to j}}\rangle}{e^{\ln\left[\frac{p(i|I)}{p(j|II)}\right]}} \right\rangle_{I\to II}$$

where $I$ and $II$ correspond to macrostates of the system, each composed of it's own microstates. $\langle e^{-\beta\Delta Q_{i\to j}}\rangle$ is the average of the exponential of the dissipation over all paths from microstate $i$ to microstate $j$. $\langle\rangle_{I\to II}$ indicates the average over all microstates of $I$ to microstates of $II$. Also $\beta = 1/k_B T$ here, and not the Lagrangian parameter from previous sections. Using the above equation we ask the following question - what can we say about a system that maintains it's macrostate $I$ when driven by an external field. If macrostates $I$ and $II$ are identical, then $\pi(I\to II) = \pi(II\to I) = 1$. For the purposes of this essay, let us assume we are dealing with a large enough state space to approximate $p(i|I) = p(j|I)$. This reduces the equation to

$$\left\langle \langle e^{-\beta \Delta Q_{i \to j}} \rangle \right\rangle_{I \to I} = 1$$

We see that the average of the exponential of dissipation over all paths from all microstates $i$ to $j$, both belonging to macrostate $I$ is equal to 1. Taking negative logarithm on both sides and using the cumulant generating function we get

$$\begin{aligned} -\ln \left\langle \langle e^{-\beta \Delta Q_{i \to j}} \rangle \right\rangle_{I \to I} &= \beta \langle \Delta Q \rangle_{I \to I} - \frac{\beta^2}{2} \sigma^2_{\Delta Q_{I \to I}} \cdots \\ &= \kappa - \gamma = 0 \end{aligned}$$

where $\kappa = \beta \langle \Delta Q \rangle_{I \to I}$ is the mean dissipation, and $\gamma = \kappa + \ln \left( \langle e^{-\beta \Delta Q_{i \to j}} \rangle \right)_{I \to I}$ represents the fluctuations about the mean. If $\kappa = \gamma$, that would mean that the fluctuations about the mean are equal to the mean. We are especially interested when $\kappa = \gamma$ is low. Since both the mean and fluctuations about the mean are low, we would have dissipation over all paths from one microstate to another, within a macrostate to be low. This is exactly the type of physical system that I would label a minimally dissipative system, expected to show learning dynamics. This is not contradictory to the dissipation driven adaptation hypothesis, where we compare $\pi(I \to II)$ and $\pi(I \to III)$. The state $II$ is more likely and seems adapted if for equal amounts of fluctuation $\gamma$, the mean dissipation $\kappa_{I \to II} > \kappa_{I \to III}$ and $\pi(I \to II) > \pi(I \to III)$. Thus the explanation of individual learning under the minimal dissipation hypothesis does not conflict with England's dissipation driven adaptation.

## V. Reasons to be Excited

In this section, I will present reasons to be excited about of the minimal dissipation hypothesis. Let ask the question, what should subsystem interactions of a system $\mathcal{S} = \mathcal{S}_0 \mathcal{S}_1$ that dissipated minimally look like? It is possible show that for $p(k^{\mathcal{S}_0}, l^{\mathcal{S}_1})$, we have

$$p(k^{\mathcal{S}_0}, l^{\mathcal{S}_1}) \propto \sum_{i^{\mathcal{R}_0}} e^{\frac{-1}{\beta} D_{KL}[p(j^{\mathcal{R}_1} | i^{\mathcal{R}_0}) | p(j^{\mathcal{R}_1} | k^{\mathcal{S}_0}, l^{\mathcal{S}_1})]}$$

The above equation indicates that the correlation between subsystems $\mathcal{S}_0$ and $\mathcal{S}_1$ changes as required to maximally predict the future inputs and achieve minimal dissipation, imbuing it with the *plastic* nature that we have come to expect in various levels of a neuronal system. Thus concepts like plasticity that are fundamental to neuroscience can be derived from this larger principle.

In an earlier section, I had discussed the connection to the IB method. This allows us to tie the hypothesis with techniques that have already shown to be related to the IB method, and make use of the vast results and insight that has been gained. These include biologically inspired signal processing techniques like slow feature analysis [15] and sparse coding [9]. The connection to other unifying princples like Friston's Free Energy principle [11] and Hawkins' Hierarchical Temporal Memory [16] can be made. The fact that the hypothesis here is physically grounded and derived from fundamental law gives it an unique advantage.

Another idea that has gained significant momentum in neuroscience recently is the *criticality hypothesis* [17]. It states that the collective dynamics of large neuronal networks in the brain operates close to the critical point of a phase transition, where the brain capacity for information processing is enhanced. According to this hypothesis, the activity of the brain would be continuously transitioning between two phases, one in which activity will rapidly reduce and die (sub-critical), and another where activity will build up and amplify over time (super-critical). The theory states that the human brain is an example of a system exhibiting self-organized criticality, with a critical

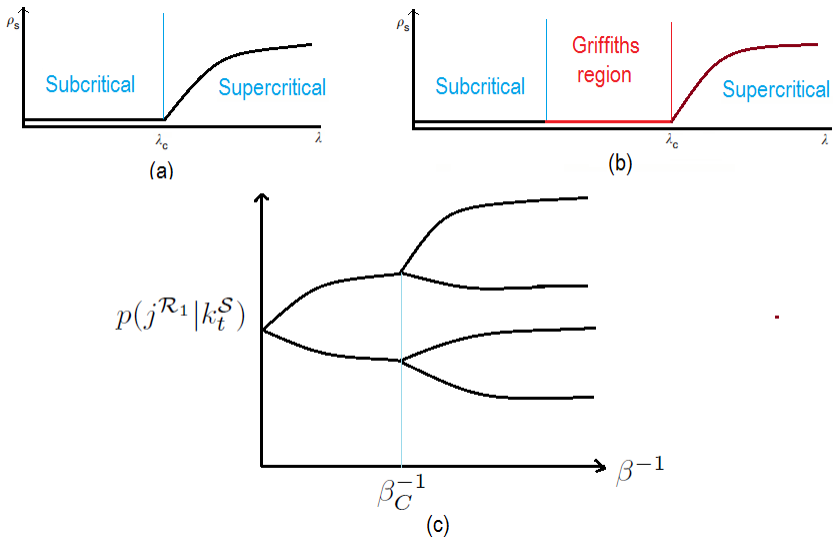


Figure 3: (a) Second-order phase transition between an active and inactive phase at critical point $\lambda_c$. (b) Phase transition with the extension of the critical point to a Griffith phase. (c) A simple figure showing bifurcations at critical $\beta_C$ in $p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})$ vs $\beta^{-1}$ graph, of a system that is minimally dissipative.

point at an attractor and a constant tuning of the control parameter by a decentralized internal mechanism. I will next show here that systems which dissipate minimally can exhibit critical behavior and establish a strong connection between the two hypotheses.

The solution we get from constrained optimization are self-consistent equations corresponding to a family of curves on the $(\Delta\langle E^{\mathcal{B}}\rangle, \mathcal{I}^{\mathcal{R}_0\mathcal{S}})$ plane, and are parameterized by $\beta$. We can use the equations for $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ and $p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})$ to form an eigenvalue problem

$$\left[I - \beta C(\beta, i^{\mathcal{R}_0})\right] \frac{d \ln p(i^{\mathcal{R}_0}|k_t^{\mathcal{S}})}{dk_t^{\mathcal{S}}} = 0$$

The critical $\beta_c$ obtained as the non-trivial eigenvalues of this eigenvalue problem are points at which the curves bifurcate through second order phase transitions (Fig.(3c)). Results for the case of Gaussian inputs (Fig.(1) in [18]) show that the system undergoes multiple continuous second-order phase transitions through many critical points $\beta_c$. We can see that the optimal input mappings of systems that are minimally dissipative thus exhibit a Griffith's region (Fig.(3b)) [17], an extension of the critical point to a larger region. If the brain can be viewed as a minimally dissipative system, it should not be surprising that it's information processing capabilities is optimal in the critical regions of $\beta$, hence explaining the properties of criticality observed in neuronal avalanches, branching parameters, etc..

We refer to Fig.(3c) here, and Fig.(1) from [18], indicating the type of bifurcations that is expected in $\mathcal{L}$ as $\beta$ is varied. The symmetry breaking at these critical points can correspond to different optimal mappings for the same value of $\beta$, allowing for different predictions by the same system. As $\beta \to \infty$, we can see that all the mappings are uniform and there is no emphasis on prediction, similar to the soft limit in a clustering problem. For $\beta \to 0$, there are no finite complexity constraints (which would be improbable), emphasis is on prediction and we are left with the equivalent of the hard clustering limit. As long as the system maintains the value of $\beta$ in the critical regions, the system will exhibit optimal information processing dynamics. If we were

to use a broad definition for cognition as the process of acquiring information and understanding it through experience and the senses, we can view the spectrum of $p(k_t^{\mathcal{S}}|i^{\mathcal{R}_0})$ and $p(j^{\mathcal{R}_1}|k_t^{\mathcal{S}})$ as different levels of cognition. We would then have the critical Griffith regions with enhanced information processing corresponding to optimal cognitive processes. This would leave us to interpret moving towards the super-critical $\beta \to \infty$ and sub-critical $\beta \to 0$ regions as those of sub-optimal prediction capability, and an impaired state of cognition. I would conjecture that both lack of any neuronal activity like in brain death, and large synchronous neuronal activity that produce epileptic seizures correspond to the ordered sub-critical stage (epileptic seizures are traditionally considered to be part of the super-critical regime). As $\beta$ is increased and we move toward the disordered super-critical region, we could obtain states corresponding to sleep, and even being unconscious.

Since an entire separate essay would be necessary to properly address the topic of consciousness, qualia and meaning, I will not get into it here. I will note while information integration, recognized as an important feature of consciousness is especially high in complex networks near criticality [19], by specifying conditions on which systems can exhibit optimal information processing dynamics, the minimal dissipation hypothesis can avoid the trap of panpsychism. This adds further confidence that minimally dissipative systems operating in the critical region are very good candidate systems for consciousness.

## VI. Conclusion

In this essay, to establish that intention is physical I used the minimal dissipation hypothesis to explain the emergence of learning dynamics and goal oriented agency in physical systems, as a trade off between energy dissipation and complexity. I also showed that there is no conflict between the minimum dissipation hypothesis and dissipation driven adaptation to unify individual learning with evolutionary processes. The connection between this hypothesis and the critical brain hypothesis looks very promising and provide a very useful approach to study questions regarding intelligence, agency and maybe even consciousness in a physically grounded manner. A multi-disciplinary approach would be needed to address questions of immediate interest, such as the phase space characterization of self-organized systems which dissipate minimally, improved understanding of internal control mechanisms to maintain criticality, and detailed formulations of cognitive states as phase transitions in a (non-chaotic strange) attractor.

## Acknowledgements

I would like to thank Prof. Neal Anderson for many insightful discussions, and all my colleagues and friends for a critical reading of the essay and the editorial suggestions. I would also like to thank the other contributors to this essay contest, whose entries have constantly influenced my own ideas and presentation (under the minimal dissipation hypothesis of course!). The future is indeed very exciting.

## References


[1] Hasenstaub, Andrea, et al. "Metabolic cost as a unifying principle governing neuronal biophysics." Proceedings of the National Academy of Sciences 107.27 (2010): 12329-12334.

[2] Landauer, Rolf. "Irreversibility and heat generation in the computing process." IBM journal of research and development 5.3 (1961): 183-191.

[3] Anderson, Neal G. "On the physical implementation of logical transformations: Generalized L-machines." Theoretical Computer Science 411.48 (2010): 4179-4199.

[4] Ganesh, Natesh, and Neal G. Anderson. "Irreversibility and dissipation in finite-state automata." Physics Letters A 377.45 (2013): 3266-3271.

[5] Still, Susanne, et al. "Thermodynamics of prediction." Physical review letters 109.12 (2012): 120604.

[6] Sartori, Pablo, et al. "Thermodynamic costs of information processing in sensory adaptation." PLoS Comput Biol 10.12 (2014): e1003974.

[7] Crutchfield, James P. "The calculi of emergence: computation, dynamics and induction." Physica D: Nonlinear Phenomena 75.1 (1994): 11-54.

[8] Tishby, Naftali, Fernando C. Pereira, and William Bialek. "The information bottleneck method." arXiv preprint physics/0004057 (2000).

[9] Olshausen, Bruno A., and David J. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." Nature 381.6583 (1996): 607.

[10] Rao, Rajesh PN, and Dana H. Ballard. "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." Nature neuroscience 2.1 (1999): 79-87.

[11] Friston, Karl, and Stefan Kiebel. "Predictive coding under the free-energy principle." Philosophical Transactions of the Royal Society B: Biological Sciences 364.1521 (2009): 1211-1221.

[12] Searle, John R. "Perceptual intentionality." Organon F 19.2 (2012): 9-22.

[13] England, Jeremy L. "Statistical physics of self-replication." The Journal of chemical physics 139.12 (2013): 09B623.

[14] England, Jeremy L. "Dissipative adaptation in driven self-assembly." Nature nanotechnology 10.11 (2015): 919-923.

[15] Creutzig, Felix, and Henning Sprekeler. "Predictive coding and the slowness principle: An information-theoretic approach." Neural Computation 20.4 (2008): 1026-1041.

[16] Hawkins, Jeff, and Sandra Blakeslee. On intelligence. Macmillan, 2007.

[17] Hesse, Janina, and Thilo Gross. "Self-organized criticality as a fundamental property of neural systems." Criticality as a signature of healthy neural systems: multi-scale experimental and computational studies (2015).

[18] Chechik, Gal, et al. "Information bottleneck for Gaussian variables." Journal of Machine Learning Research 6.Jan (2005): 165-188.

[19] Arsiwalla, Xerxes D., and Paul FMJ Verschure. "High integrated information in complex networks near criticality." International Conference on Artificial Neural Networks. Springer International Publishing, 2016.