Integrating deep learning and active inference within artificial general intelligence

My Research Goals:

The Human Brain – the most fascinating computing device to exist – is becoming more popular than ever before in modern artificial intelligence; specifically, characterizing its functional architecture to establish better models of neural computation that mimic learning and inference in humans. My research hopes to progress this idea to new levels under 3 integrative approaches:

1) Understanding deep convolutional network architectures in relation to the structure to the visual cortex of the human brain to see if something like deep CNN's is present in the brain, e.g., the notion of receptive fields and the hierarchical structure used for processing **[29]**.
2) Understand the importance of attention in human perception, to create better attention modules for visual and language tasks **[3][9][14]**.
3) Combine the ideas above to create a learning system that not only learns to infer but also has some innate mechanisms to incorporate the "wiring rules" that govern the learning processes in the human brain, i.e., structure learning. This structured learning may transcend the lifespan of a human to inculcate wiring rules encoded in the human genome, which have evolved over generations **[11]**.

Main Objectives/questions to solve:

To accomplish these goals, I have based my research on solving two key problems that limits us in realizing the true potential of Artificial Intelligence:

A) Creating truly Intelligent systems, i.e., Artificial General Intelligence by working towards a more quantitative and statistical approach to consciousness and unconscious inference, the human mind and how it makes sense of the (prosocial) world **[6]**.
B) The second problem involves optimizing the existing structure of the neural networks in such a way that makes them more efficient and thereby helps them to learn better data representations. This would involve improving upon the existing methods of self-supervised/unsupervised learning to include complexity as an explicit part of the objective function – which would involve finessing the basic structure and functioning of an artificial neural network.

The Bayesian Brain:

When Helmholtz was a child, in Potsdam, he walked past a church and saw tiny figures standing in the the belfry; he thought they were dolls, and asked his mother to reach up and get them for him: he did not yet understand the the concept of distance, and how it made things look smaller. When he was older, his brain incorporated that knowledge into its unconscious understanding of the the world—into a set of expectations, or "priors," distilled from its experience—an understanding so basic that it became a lens through which he couldn't help but see which in my belief laid the foundation for a generation of researchers to finally formulate a bayesian brain hypothesis which include people like yourself, Dr. Hinton and Dr. Friston among many uncountable others.
The bayesian approach not only helps us formulate the multi-functioning and multi-dimensional brain but also highlights the various shortcomings and failures of the zenith of evolution process that our brains sits on as formulated by Dr. Kahneman and Treversky in their famous book "Thinking fast and slow".

A)Free energy and AGI (Artificial General Intelligence):

The free-energy principle [8][11][15][18] has ineluctably been the most fascinating theory for me in recent years, that tries to unify the various working principles of the human brain – and cognition – into one simple imperative for human perception, processing, action, and the integration of information that ultimately forms the basis of sentience and, perhaps, consciousness. The thing that makes free energy such an intriguing concept lies in its simplicity. The simple notion of minimizing "surprise" in order to ensure "Survival" echoes Einstein's observation that (using his general theory of relativity) no "matter" wants to disappear and every bit of "matter" moves to the point in space where time passes most slowly. The other intriguing aspect of the theory is how we can practically use the simple theoretical construct of minimizing surprise by taking in the sensory information, processing it and then modifying beliefs via changing the “internal state” of a system; ultimately resulting in an action which ensures that surprise is minimized. Put simply, the brain tries to find the best probabilistic beliefs that maximize the likelihood of observed phenomenon; hence reducing “Surprise”. Just using this simple construct – based upon Bayesian inference – can be used to construct sophisticated or deep (hierarchical) systems that take in information across multiple sensors to produce an evidence-based rational decision. Formally, every system can now be decomposed into a recognition density based upon prior evidence – encoded by its internal “state” – that generates action on the environment supplying sensory evidence. So, essentially, we have all the tools necessary to describe the process of how a single neuron – or neural network – might be processing information and hence we just now need to scale this system up to the functioning of trillions of neurons. Here, the interesting concept of Ensembles [9][10] comes into play which – in my account – underwrites system functionality and the process of evolution.

Active Inference, Deep Learning, and Similarities: I am intrigued by the ideas that the Free energy principle offers. And, in my thought experiments, I can see it is able to answer many questions; e.g., why psychopaths are psychopaths because their brain simply does not process the sensory inputs as our brains do, hence it is not able to recognize or represent love empathy, sympathy, and pain – explanations or hypotheses that would otherwise provide good explanations for other people's behavior. The challenge for me is how to use advances in deep learning and the free energy principle to create Intelligent systems, i.e., AGI Systems [19]? Why I want to combine these approaches is due primarily to the similarities in the fundamental architecture, inspired to some extent by the human brain [15].

Let me start with ANN’s and deep learning. If one considers the attentional aspect of human learning it is fascinating and frustrating at the same time because the way our brains process information – which we perceive to be with great detail when on the contrary we consciously are dealing with a minuscule amount of information – ultimately depends upon the person doing the “Learning”. In a recent paper "Ventromedial prefrontal cortex compression during concept learning" [16] it is suggested that PFC, is responsible for goal-driven learning while filtering out irrelevant information. This is McCain to dimensionality reduction and – for me – evidence for the compression phase of neural networks as articulated using the Information Bottleneck theory [13]: in this formulation, ANNs learn abstract representations from data in two-stages: the first stage is the feeding of the data to the network and calculating the loss. The second and more interesting phase is the compression of these representations in a layer-by-layer "reduction" of the data manifold, which ultimately enables the learning of abstract representations. Formally, this can be expressed in terms of free energy minimization, via minimization of complexity, i.e., Accuracy - Complexity = F, where we argmax F. Here complexity represents the divergence between prior and posterior representations: i.e., how much you need to move away from your prior beliefs to explain observed data or evidence accurately. This is formally similar to dimension reduction – to find a simpler manifold that explains our data [15][18][19]). Arguably, this is what gives ANN's such remarkable power; sometimes, to perform better than humans in certain goal-specific tasks.

The way I like to think about the dynamics of the working of neural networks is, how two different algorithmic paradigms are so closely coupled to each other i.e the divide and conquer approach and the greedy approach, the two are highly coupled because if you see by creating a layered architecture we are basically dividing your problem and then while backpropagating you are doing a layer-by-layer compression of the representation which actually is a greedy approach(i.e how you calculate the rate of change of loss with respect to weights of that particular layer, hence essentially forming one part of the chain which has as many parts as the number of layers in your network [1]). So, essentially both the techniques are similar in the way that they try to learn the data representation and the data manifold that might exist in some n-dimensional hyperspace but what differentiates the two is the way they do it, Deep learning employs a pure data approach and learns purely the “representation of data that it has been fed with” while on the contrary through active inference and generative modeling we are able to learn “probabilistic data representation” a kind of a probabilistic data manifold [1][13][19][18] . So, what if we are able to combine the two? In my mind, this may take us closer to the question of what defines human sentience [17]?

B)The Neuron, threshold and the activation function[2]: An intriguing aspect of human learning and inference – that interests me – is how do neurons fire: how do they get activated: what kind of activation do they use, and what is the threshold that it uses? Is the activation function itself learned over time? One thing I am confident about is that the activation functions in our brain are fundamentally different from those used in ANN's – and, in particular, the uniform thresholding used in this setting. Our construal of the brain is circumscribed and the systems that we call AI-based systems are even simpler. For example in **[2],** it is suggested that our brains learn different threshold values for every neuron (more than 100 billion) and the nonlinearity used is formally different from what we see in ANN's. I have been working on understanding the activation function and firing potential of neurons (w.r.t ANN's) and my hypothesis is that the activation function should be learned for every layer.

So where are we heading?

Neural Networks and evolution (2012-present): One thing that has been a prevalent part of the evolution of neural networks has been the incremental inventions that have revolutionized their applications. For example: 1) Introduction of 'relu' for which Geoff Hinton himself claimed that "we were imbecilic people to utilize sigmoid for 30 years, without genuinely understanding the fact that the function itself is saturating so will be its gradient2 **[5]**. 2) Solving the 'dead relu' quandary, give the flat line a slope, and it worked. 3) Still vanishing gradients were a resistant problem and hence Batch Normalization came in: just normalize the outputs after every layer to stop zigzagging during convergence. 4) Still training CNN was hard so emerged skip connections – just skip a few layers and integrate the residue, which introduced the concept of residual networks. 5) Now with deep networks came the problem of cumbersome computation especially with CNN's so how to downsample the feature maps without losing much information. The fix was 1x1 Convolutions or simply a layer of neurons (just for the analogy purpose).

So, in a way, these minute inventions have genuinely revolutionized deep neural nets and their applications, but with all this said, I believe we still have not genuinely understood neural networks and there are many things to explore. For example, the best paper award in NIPS 2018 was given to the paper on Neural Ordinary Differential Equations **[6]**, a paper that I thought held the key for revolutionizing problems based on time series forecasting. The authors abstracted the discrete hidden layers and parameterized the "derivative" of the hidden state with a neural network, whose output they calculate utilizing a black-box differential equation solver; i.e., they tried to find the function by integrating the derivative – and thus were able to extrapolate missing points utilizing the approximated function. We know there have been many years of research in solving ODE's. So, a plethora of methods to solve those equations – and hence a plethora of landscapes to explore in Neural networks based on ODE's arises (and this is just one example). Why do I put so much emphasis on neural networks? The answer is simple: neural networks are a set of tools that can be used to learn functions or approximate them to such precision. However, their progression depends upon replacing a reliance on incremental and *ad hoc* inventions with a first principle approach – such as the free energy principle.

Possible Approach and methodology(for part A) :

So, the issue is how and what are the possible approaches that can help combine the two? The answer is simple and complicated: simple because the most ubiquitous way to combine the two is using generative modeling and reinforcement learning approaches as recent papers have shown **[20]**. I will discuss this paper later. The first thing is to understand what the free energy principle offers as an idea and how the accompanying active inference can help us improve and leverage generative world models. The first assumption is the existence of markov blankets separating every state of the system under study from the macro to the micro level i.e. from single cells to the brain and the whole of the human body which gives rise to the process of embodied cognition.

Let us assume that the brain is a generative model that takes in sensory information as a set of observatory inputs **[22]**. Using these inputs our brain can generate perceptions, i.e., experience something without actually doing it (like the image of a blue sky over a desert landscape). Let us assume our brain is an agent with a generative model $p_\theta(o, s)$ of sensory observations o and latent variables or states s, which we can factorize into a likelihood function $p_\theta(o|s)$ and a prior on the states of the generative model (our agent) $p_\theta(s)$ as:

$$p_\theta(o, s) = p_\theta(o|s)p_\theta(s) \qquad \textbf{(equ. 1)}$$

Here θ is the set of parameters that “the brain” can update or change to update its generative model of the world. These parameters can be thought of as the weights between neurons. Now the Integral that attends **equ. 1** turns out to be very difficult to solve – as the agent would be required to marginalize over all the possible states for a particular observation. Therefore, variational procedures employ a simple trick, i.e., rather than assuming the agent directly minimizes surprise $-\ln p_\theta(o)$ directly, assume rather it minimizes an upper bound, which is a lot simpler to calculate.
Using the fact that the Kullback-Leibler (KL) Divergence:

$$D_{\mathrm{KL}}(p_a(x)||p_b(x)) = \int_{x\in X} \ln\left(\frac{p_a(x)}{p_b(x)}\right) p_a(x)\,\mathrm{d}x .$$ **(equ. 2**)

between two arbitrary distributions $p_a(x)$ and $p_b(x)$ with shared support on a common space X is always greater than or equal to zero, and equal to zero if and only if the two distributions are equal (**Notion 1**). The way this notion helps is very similar to working of the Loss function in neural networks, i.e., now minimizing free energy simply becomes minimizing prediction error. Using this simple notion, we can define the free energy as:

$$F(o,\theta,u) = -\ln p_\theta(o) + D_{\mathrm{KL}}(q_u(s)||p(s|o)) \geq -\ln p_\theta(o)$$ (**equ 3**).

Here $q_u(s)$ is recognition or variational density over the space of latent states s, which belongs to a family of distributions parameterized by a time-dependent, i.e. fast-changing, parameter set u. Here u is actually $u = \{\mu, \sigma\}$, represents the agent’s recognition density q_u(s) and p(s|o) is the true posterior. Thus optimizing $u$corresponds to minimizing the Kullback-Leibler divergence between the variational distribution $q_u(s)$ and the true posterior distribution $p_\theta(s|o)$. Thus, the agent is automatically able to represent a probabilistic representation of an approximate posterior on the states of the world, given its sensory input. Now, as the variational free energy upper bounds the surprise $-p_\theta(o)$, minimizing the free energy w.r.t the parameters theta of our generative model would simultaneously correspond to maximizing the evidence $p_\theta(o)$ for the agent’s generative model of the world. This results in the optimization of our generative model w.r.t model parameters, this is also termed as perceptual learning.

Now comes the interesting part “the active part”: till now we have seen that the agent can minimize the free energy bound by learning (optimizing the parameters of its generative model) and perception (optimizing the sufficient statistics of the variational density) but we can also change how or what observation we actually make; i.e., we change the observations itself, which then changes both our learning and perception. This is called **Active Inference**. So essentially, a theory of how humans learn and perceive suggests that both learning, and perception are intimately related and provide a first principle account of how we sample the world. Such a lucid theoretical explanation has made me admire this formulation. Finally, this theoretical approach gives us the following dynamics for the parameters θ of an agent’s generative model, its internal states u, encoding the sufficient statistics of the variational density q, and the states of its actuators(actuators refers to our arms or muscles through which we interact with the environment) a:

$$(\theta, u, a) = \underset{(\theta^*, u^*, a^*)}{\arg\min} F(o(a^*), \theta^*, u^*)$$ (**equ 4**).

In principle, this covers (nearly) everything: we can create an agent for a goal-specific behavior by encoding specific goals in the priors of the generative model: assigning a higher prior probability to a specific state, means that the agent will try to seek out this state more often. Here is where a plethora of opportunities lie – as we already know updating the parameters of our agent would require updating over slow time scales, which would mean large batches of data and the internal states u and the action states a must change on the timescale of the sensory input. Thus, in a discrete world, they would have to be optimized at each time step, which would be computationally burdensome.

It is at this point we can use the universal approximation power of deep neural nets; namely, we can optimize the parameters together with the parameters of the generative model (there are a few existing works on this which include **[21]**). This amortization (i.e., learning to infer) is what was achieved by the author of the above-mentioned paper **[20],** where the author learned a bootstrapped estimate of the density function from samples using a neural network to learn an approximate inference distribution. He although used an approximate expected-free-energy (EFE)-value network but the notion remains the same. I acknowledge that the premises and consequences of active inference requires a significant amount of investment. However, if one gets used to its notions and can understand what it really conveys – and how it formulates and embeds in itself the process of learning, perception, and action – one can absorb the process of human learning under a simplistic notion of minimizing prediction errors. In turn, one can inflate this simplistic notion within a complex Bayesian framework that ultimately gives it the power to explain very complicated tasks. From optical illusions (e.g., the famous moon illusion) to neural responses to oddball stimuli, mirror neurons, predictive coding, basic features of neural architecture, critical neural dynamics, and human choice behavior. Many empirical findings can be framed and explained by the free energy principle and the resulting active inference.

A Philosophical take on the evolution of Human Intelligence:

Humans, Humanity, Mankind, all sound kindred. They are though quite similar and define the same thing i.e "US"(Not USA) and our struggles since we came into existence from the Paleolithic era to the Neolithic era. Human history can be broadly relegated into four revolutions as per Yuvai Nova Harari: 1)The "Cognitive Revolution" 2)The "Agricultural Revolution 3)The Unification of Mankind(this is where it all commenced the elevate of ecumenical imperia and Capitalism that is marring our societies today) and last but not the least, as it is always the case "The Scientific Revolution" which today has brought mankind on the verge of a revolution that will shape our world and the definition of "Humanity" for the future generations to come. Whatsoever be the period or the age one thing that has been with us(Since the commencement) and is a component of all of us right now is Human "greed", whether it be the fall of imperia in the past to fall of societies and cultures today or to the invention of the wheel or to our struggle to engender Artificial General Intelligence, it is a thing that we all are born with and the only distinction is some are acquisitive for others sake while some are for themselves, I like to think I fall in the prior category but it is what has driven our curiosity, it is what makes us "Humans" and differentiates us from being animals it is because of this we persist to struggle for the things we love or we are endeavoring to protect. The endeavor to understand the "Human Brain" and to unlock the true nature of consciousness lies at the very core of raveling the true nature of Human Intelligence, the brain, as astounding are its capabilities even more astounding is the intricate structure and the potency it gives us, the humans to relate their past with present and take actions in order to ameliorate their future, man is the only living thing that is able to comprehend past, present, and future into one being and this is what sets us apart amongst all the living things to have ever grazed this planet.

My instincts give me the intuition that " By looking at the brain one can essentially tell if that brain belongs to your universe or it is a brain from another universe". This made me agree with the fact that the hierarchical structure of the brain in itself embodies every aspect of the environment that surrounds that 'system' in which that brain exists. I came to the same conclusion around 6 months back when I was trying to solve my most difficult set of problems. The problem that I was trying to understand back then is established on two fundamental questions: 1) What would have been Prof. Hinton's and Dr. Rumelhart's inner thoughts that led them to derive the backpropagation algorithm that we use today to train the deep neural networks(although Backprop is an old idea dating back to 1960s), the bigger part of this question involved the logic that they used to optimize a complex network involving so many parameters and they addressed the problem using a very minimalistic approach, a simple gradient calculation using chain rule for every level of composition and what amazes me is the intuition behind such formulation, so I asked myself does the brain do backpropagation or it is simply a mathematical approach to optimize composite networks? The second question is much broader in the sense that it involves interdisciplinary expertise, and this question is one that I have struggled for very long to even start answering it in the first place. The question is "what we as humans do, why do we do it? Why do the irrationalities in our mind take control of the rational being in us resulting in negligence of every possible evidence that can help that system to modify its recognition density".And now the main problem that engulfs both the above-stated questions is, can we really treat the brain simply as a Bayesian inference engine or there is more to the story than just a generative probabilistic system working in such a way to reduce surprise and to ensure his/her genes are passed onto the next generation? The question itself I think takes us to another deeper question and the nature of human consciousness-expanding itself into the deeper debate of Nature Vs Nurture. The Brain that we all are today fascinated with, Is it a result of nature evolving us in such a way to ensure the existence of human civilization, or the brain itself has created a society that is as complex as the neural wirings of the circuits in our brain that makes it ever so complex to even comprehend its true nature.

What really fascinates me is the multitude of the various dimensions of human nature and its form of expression, the possibility of everything and nothing that our brains are capable of achieving is what really begs me to ask what it is that defines the true nature of humans and I feel history has given some of the answers to this question, starting from Dr. Edward Bernays who truly understood the power of the irrational mind that lies just beneath the surface of the modern scientific civilized society and how it can be used to manipulate the masses on the wishes of a few handfuls of people. If you go back in history, even more, Colonization i guess is the biggest example in human history of how mass oppression is done in a regularized and a structural form where fellow human beings are not considered even humans in the first place. So what kind of sensory inputs lead to such modification of the recognition density which leads to a generative model that itself is trying very heavily to dominate the other generative models in the environment, which in turn is able to some extent answer the question of what happens when 600 billion brains try to exist together, an ensemble of systems where every system is more complex than anything we have ever seen as a civilization and when such individually complex systems are confined on a terrestrial planet left to interact with each other leads to the current society that we have, so what has shaped us, is it the nature of our brains and the biology behind it or the system that comes into existence when complex systems try to work together(Nurture) or it is a mix of both or it is something about the human mind that we haven't yet seen or observed.

Conclusion:

As it stands the Free-energy principle, theory or a framework whatsoever one may call it, at its very core is not falsifiable at least that's what I have understood because of the vast space of priors and posteriors that it encapsulates within itself, an Idea that tries to approximate a distribution with another distribution can be called a variational trick as it is used in Physics to calculate intractable partition functions with infinitely many states but this simple trick might really reveal the truism behind the ever evolving human brain. I believe with the free energy principle and the corresponding active inference framework as proposed by Dr. Friston can actually turn neural nets and the current field of AI into something close to the general-purpose human thinking machine i.e. The brain.(Maybe at the very least it might just reduce the 500 year timeline by half).

References:

[1]: Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. *Learning internal representations by error propagation*. No. ICS-8506. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[2]: Sardi, Shira, et al. "New types of experiments reveal that a neuron functions as multiple independent threshold units." *Scientific Reports* 7.1 (2017): 18036.

[3]: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[4]: Bello, Irwan, et al. "Attention augmented convolutional networks." *arXiv preprint arXiv:1904.09925* (2019).

[5]: Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[6]: Chen, Tian Qi, et al. "Neural ordinary differential equations." *Advances in neural information processing systems*. 2018.

[7]: Linear Algebra by Prof. Gilbert Strang fall 2019, MIT OpenCourseWare.

[8]: Hierarchical models in the brain, K Friston - PLoS computational biology, 2008

[9]: Learning and inference in the brain, K Friston - Neural Networks, 2003

[10]: Free-energy and the brain, KJ Friston, KE Stephan - Synthese, 2007

[11]: Yamins, Daniel LK, et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111.23 (2014): 8619-8624.

[12]: A hierarchy of time-scales and the brain, SJ Kiebel, J Daunizeau, KJ Friston - PLoS computational biology, 2008

[13]: On the information bottleneck theory of deep learning, AM Saxe, Y Bansal, J Dapello, M Advani, A Kolchinsky… - Journal of Statistical Mechanics: Theory and, 2019.

[14]: The functional anatomy of attention to visual motion. A functional MRI study., C Büchel, O Josephs, G Rees, R Turner, CD Frith… - Brain: a journal of neurology, 1998.

[15]: Hobson, J. Allan, KJ Friston. "Consciousness, dreams, and inference: The Cartesian theatre revisited." Journal of Consciousness Studies 21.1-2 (2014): 6-32.

[16]: Mack, Michael L., Alison R. Preston, and Bradley C. Love. "Ventromedial prefrontal cortex compression during concept learning." Nature Communications 11.1 (2020): 1-11.

[17]: Bourdieu, Pierre. "The forms of capital." (1986): 258.

[18]: Variational free energy and the Laplace approximation, K Friston, J Mattout, N Trujillo-Barreto, J Ashburne - Neuroimage, 2007.
[19]: The free-energy principle: a unified brain theory? By KJ Friston.
[20]: Deep Active Inference as Variational Policy Gradients, Beren Millidge.
[21]: Scaling active inference, Alexander Tschantz, Manuel Baltieri, Anil. K. Seth, Christopher L. Buckley
[22]: Stochastic Backpropagation and Approximate Inference in Deep Generative Models, Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra
[23]: Auto-Encoding Variational Bayes, Diederik P Kingma, Max Welling.
[24]: Deep Active Inference, Kai Ueltzhöffer.
[25]: Spiking neuron models: Single neurons, populations, plasticity, W Gerstner, WM Kistler - 2002
[26]: On the choice of metric in gradient-based theories of brain function, SC Surace, JP Pfister, W Gerstner, J Brea - PLOS Computational Biology, 2020
[27]: Biologically plausible deep learning—But how far can we go with shallow networks?, B Illing, W Gerstner, J Brea - Neural Networks, 2019
[28]: An Approximate Bayesian Approach to Surprise-Based Learning