 doi:10.20944/preprints201908.0318.v1

Epistemological and Ethical Implications of the Free Energy Principle

Eray Özkural
Celestial Intellect Cybernetics
celestialintellect.com

August 29, 2019

Abstract

The free energy principle states that self-organization occurs through minimization of free energy, which is a measure of potential thermodynamic work. By minimizing free energy, the organism happens to also minimize surprise over its boundary, promoting chances of survival. We discuss the ethical implications of the cognitive goal in detail from an empirical point of view, highlighting the principle of least action as a physical basis of Occam's razor, the universality of the free energy principle, and its explanation of natural selection. We explain that the free energy principle extends to groups of organisms and helps us understand group-scale adaptations and selection in biology. The free energy principle applies to all scales of organization in the organism from single cells to the entire nervous system. When this principle is taken to its logical extremes of modeling groups, populations and ecosystems, we uncover a new, evolutionarily sensible path at explaining puzzling aspects of human motivation and judgement, including ethical decisions. To minimize free energy, populations have to act to maximize gathering of information, while building effective models at mitigating changes to its dynamic structure. The free energy principle thus provides a naturalistic explanation of some of our deepest ethical intuitions, and valuable principles of social behavior. We interpret the cognitive goal that corresponds to the principle as seeking a dynamic, fruitful, yet peaceful activity that sustains the organism. This state of mind is interestingly similar to the Buddhist intuition of mental equanimity; the organism's final goal is to be at peace and harmony with the environment. Another immediately relevant aspect is that assemblies must form to promote symbiotic, synergistic, positive feedback loops, which coincides with the findings of ecologists. Therefore,

 doi:10.20944/preprints201908.0318.v1

ethics naturally emerges in self-organizing systems. Assemblies of organisms must ultimately unite in macro-minds to achieve the greatest reduction in free energy, as well as building technological extensions of themselves to improve their capacity to do such, therefore the principle also predicts a post-singularity world-mind composed mostly of artificial intelligence.

Keywords: AI; ethics; safety; autonomy; free energy principle; reductionism; symbiosis

1 Introduction

Without delving into the mathematical formalism, the free energy principle may be stated as the following: to survive, organisms must minimize the thermodynamic free energy in a biological system. The free energy, or Gibbs energy in a chemical system is a measure of how much thermodynamical work the system is capable of performing. Above that limit, the system will not participate in chemical reactions. The more free energy an organism has, the more reactions may happen due to chance. This leads to disorder, and the more disorder a system suffers the more chance there is for its structure to dissolve. The organism must therefore minimize surprising interactions with the environment to maintain integrity [10]. The integrity condition has been elegantly defined as a boundary called the Markov blanket that discriminates the organism from the rest of the environment, decoupling internal states from external states.

Although classical understanding of potential energy is usually applicable to systems at equilibrium, as in classical chemistry, the Gibbs energy applies to all physical systems, that is to say it works for dynamical systems far from equilibrium. The formalism of free energy rests on a stochastic dynamical model of a phase space (any physical system), and therefore the interactions are described probabilistically. The Markov blanket thus is defined as an insulation of the organism from external effects, only mediated through the blanket via special states we may recognize as sensory and active states. The cell membrane is a good example of the Markov blanket as it neatly partitions the environment into internal organism, and external environment states as an insulation layer. The environment inside the blanket is controlled by the cell, nutrients and sensory data are obtained through the membrane, and the cell may have active states such as those that modulate the membrane, or flagella. Likewise, the nervous system in an animal may be considered to have a Markov blanket, wherein the mediation with the environment occurs through stimulation of receptor and effector neurons. The brain has to continually infer causes in the environment, and predict future states to choose the actions that will minimize free energy, and survive. In the free energy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2019 doi:10.20944/preprints201908.0318.v1

Energy Principle - Classical.pdf

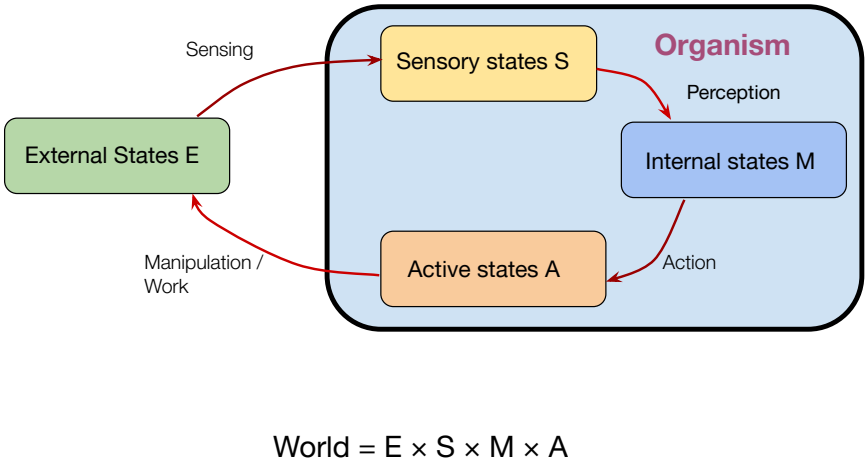

Figure 1: The cognitive model implied by the free energy principle, sometimes termed the active inference agent or the autodidactic agent. The world is partitioned into environmental, sensory, internal, and active states.

formalism, all of these state sets and possible interactions among them are described as probabilistic, dynamical interactions affixed in physical space. The entire physical system evolves, stochastically, and so do the organism's various components such as its internal states, sensory states, and active states. Fig. 1 depicts the cognitive organism model implied by the free energy principle, and the stochastic interactions among the state set partition of the world according to the model.

One of the foundational results of the free energy approach is that minimizing free energy entails minimizing surprise, relative to the organism. Surprise is equivalent to expected entropy; if my organism's boundary has to face so many random interactions over a time window, sooner or later my boundary will be punctured, and my organism will disintegrate. That is why, in nature, the unknown and the complex elicits so much fear to animals; as a dog barks at the storm or the vacuum cleaner, this is because it cannot reliably predict what sorts of random interactions will be caused by

 doi:10.20944/preprints201908.0318.v1

it. This equivalence is the defensive, or protective condition of life, and at the shortest time scale it manifests as defensive reflexes. However, to enable long-term survival, a complex organism cannot rely on fixed rules. Instead, it must continually discover and learn to model the environment, so that the survival condition, including feeding and healing, may be met through actions that manipulate the environment (and the organism) to the organism's advantage.

Even in a cell, there is complex signal processing, called the signaling network, as opposed to the metabolic network, that processes information in non-trivial ways. Although not quite intelligent, bacteria can adapt to nutrient gradients, and that is a sort of adaptive behavior. This adaptive behavior is not merely a curiosity, it is enforced by the free energy principle. To live that long, the organisms had to evolve adaptive behavior. There are two main sorts of cognitive adaptations that the organism can follow to minimize its free energy. Firstly, it may optimise the accuracy of its world model, i.e., adjust its internal states to better match the environment, and therefore infer past and future causal states more effectively, enhancing all manners of thought about the world. Secondly, it can creatively manipulate its environment to confer a survival advantage, as predicted by Chomsky when he pointed out a serious shortcoming of behaviorism: a behaviorist mind is not inclined to improve over existing natural conditions [4].

Friston and his research team model the learning process by Bayesian inference. The organism updates probability distributions corresponding to a model of the sensorium. The organism then seeks to minimize the difference (called Kullback-Leiber or KL divergence) between the recognition (a model of environmental states) and generative (a model that will generate environmental states) densities. The second foundational result is that minimizing free energy entails minimizing KL divergence. If the KL divergence is zero – in which case we would have a perfectly accurate predictive model – then free energy is just the entropy of the sensorium. In reality, of course, no organism can achieve such a state, for the universe is in constant flux. At the very least, the internal dynamics requires interaction to sustain its course. Therefore, the kind of prediction here is not merely modelling the immediate environment, which might be an overtly simple environmental niche, but it has to be a world-model, encompassing the world at large. There, we glimpse at the philosophical repercussions of the principle.

And in a rather surprising way, as a third major result, we find that the free energy principle is also directly equivalent to the principle of least action. This equivalence was first observed by Feynman as he probed the foundations of quantum mechanics. The principle of least action states that a physical system will choose the most *economical* way among many possible ways, and

it turns out that this is equivalent to Bayesian inference, which in turn is equivalent to the free energy principle. Then, the principle is not merely encoded in the accidental mechanics of evolving organisms, but also nature itself. It also seems circular, for if free energy principle is provided by physics itself, does not this mean that life is necessarily caused by physical law; is not this against evolutionary science, or is not this a vicious circle? It turns out that there is no real deductive loop, or assuming the conclusion here. What this result truly means is that the Cosmos is inclined to evolve from a non-biological, unorganized state with no long-lasting Markov blankets, to biological organisms, explaining *biogenesis*. Then, we are not merely looking at a theory of neurological organization or biological principles, but a unified, cosmic theory of evolution. After all, an open-ended stochastic dynamical system like the Cosmos, evolves through time, attractor sets form, and the system continually explores new behaviors. Indeed, there is no unwarranted assumption, or circularity in this manner of theorizing. What perplexes the common reader is the universality of the theory.

Many conceptual objections have been raised against the free energy principle. Most common objections depend on the intuition that such a simple mathematical explanation is too good to be true; how can the definition of life be crammed into a few dynamical system equations? Or alternatively, critics have claimed that the framework is a sort of mathematical babble that obscures any scientific findings, and nobody really understands it. The present article is, at least partially, an attempt to address these philosophical and conceptual objections, since it does not appear that the findings require any mathematics to understand them, or failing that, we hope that we can explain them satisfactorily in usual philosophical parlance.

As we have seen, the free energy framework does contain some counter-intuitive conclusions, and is thus it is understandable why there is much critique. When viewed objectively, however, the proofs are easy to check, and there are only a few general, but weak assumptions such as ergodicity, and the work is supported by many diverse empirical studies. The objection from obscurity is not valid, if anything, the theory is too simple, and must be refined with more detailed models to apply to the real world in many cases, such as the hierarchical, predictive coding, message passing model of the brain [9]. This is akin to the case with universal AI models, where we can perhaps express a universal utility maximization agent such as AIXI [14] succinctly, but to realize it many detailed models, architectures, and methods are required. In turn, the most persuasive empirical evidence of the free energy principle will be to build an autonomous AI agent based on these principles, and test the predictions of the theory, such as, "will it tend to explore the environment?", "will it show curiosity?", "will it form habits?",

 doi:10.20944/preprints201908.0318.v1

and so forth. Friston et al. detail an autonomous agent model in [11]. Please see [8] for a self-contained review from the neuroscience perspective.

2 Epistemological Implications

From a philosophy of science perspective, there are two pertinent implications of the free energy principle that may not immediately appear to scientists. First, the free energy principle effectively reduces to the principle of least action, which is a common property of quantum mechanics and general relativity. Therefore, as far as our present knowledge of theoretical physics goes, the principle states that life, cognition and evolution, directly reduce to physics. This is a bold claim that is underemphasized in the scientific articles because physical scientists are already reductive physicalists. However, until the free energy principle, no such comprehensive theory existed that explains both a general physical basis, and dynamics. From an AI theory viewpoint, the principle of least action is related to our previous work reformulating algorithmic information theory in terms of physics [20]. Hutter et al. previously explained that the universal induction theory works, as it acts as a formalization of Occam's razor [17, 24]. However, the physical formulation in [20] goes deeper, as we introduced minimum volume, minimum energy, and minimum (quantum) action complexities. If the main claims of the free energy principle are correct, then there is a *physical* reason why Occam's razor is true, and it is highly likely that the correct universal inductive bias is that of least (quantum) action.

The second implication is that the free energy principle is a lower-level, or more general theory than whatever physical law we discover in the future, as long as it can be formulated in the stochastic dynamical system formalism. That is quite interesting, as it does not apply to only our universe, but assuming a multiverse, it would apply even to other universes in the ensemble. Astrophysics does model the evolution of complex structures such as planetary systems, however, the free energy principle might explain why cosmic evolution occurs. That is yet lower level than astrophysical models. Furthermore, the free energy principle can explain the emergence of any stable structures, which implies that it might yield a unified theory of physics. This property is intuitively unbelievable, as we have been trained by particle physicists to equivocate foundational physical theory with the standard model, which is not necessarily the most elegant theory in physics. The standard model advocates routinely dismiss even string theorists who proposed a more elegant theory with higher dimensionality – Witten has famously proposed M-theory with 11 physical dimensions. If we can explain the particle

 doi:10.20944/preprints201908.0318.v1

zoo of the standard model with the free energy principle, of course, it would be a radical success of the theory, however, this need not be true for it to achieve a feat of equal epistemological significance.

The free energy principle gives us a new tool to understand *natural selection* in terms of physics. Similar popular accounts have existed for long. In science fiction literature, Asimov famously analyzed the problem of defeating the second law of thermodynamics, in his story "The Last Question". This physical approach to life was sometimes called extropy, because it looks like it amounts to exporting entropy from the organism, or achieving negative entropy, countering the second law of thermodynamics, order against chaos. The free energy model shows that defeating the second law is likely in the realm of far-from-equilibrium dynamics. Although, we cannot guarantee evading the heat death of the universe like in Asimov's story, we might be able to explain why life exists at all for we now have a reductive definition of life, and potentially a complete understanding of universal fitness. The free energy principle thus yields a theory of Cosmic Evolution surpassing the restricted laws of Genetic Evolution, easily explaining such biological phenomena like epigenetics, and horizontal gene transfer in bacteria, but also the most foundational biological problem of biogenesis. The explanation of biogenesis is mostly opaque from the articles on biology and neuroscience, however it does exist. First, the theory does explain inorganic self-organization such as crystal growth, and sycnchronized pendulums, just as well as active inference agents [15], this is essential to explain how the various cellular machinery evolved prior to the cell. And secondly, the theory predicts that the world will follow trajectories wherever higher plateau of fitness are explored, which seems to give us an *arrow of evolution*. The principle does not necessarily lead to higher structural complexity, but it does tend to selection of adaptive structures with superior minimization of free energy,

3 Scale-Free Property and Group Organization

Another striking property of the free energy principle is that it applies at all scales of organization from the nucleotide sequences to organelles, to cells, organs, the brain and the entire organism. If that is true, it surely does not stop there, and applies to interacting organisms, families, groups, tribes, ecosystems, the biosphere, and beyond. As the scale grows, it becomes increasingly harder to visualize how the principle would apply, therefore a cogent explication is in order.

 doi:10.20944/preprints201908.0318.v1

One of the most intriguing debates in evolutionary biology is that of group or kin selection. Clearly, there are such dynamics, yet some biologists such as Dawkins insist that the individual organism unit is where any natural selection applies. This contentious hypothesis has not however been analyzed in much depth. The selection applies at the level of the individual specimen, truly, however, the individual carries genes that adapt it to groups. Traits such as monogamy or mating frequency are dictated by genes in nature, however, there is a lot more to group dynamics than merely family scale organization. A popular kind of group study in zoology is that of the extended wolf family or tribe, which is called a wolf pack. There are however social interactions in animals from such family like units to large tribes. All primates, for instance, show a manner of tribal collective behavior, but so do many other familia, such as flocks of birds, herds of herbivores, and ants and bees. These tribal behaviors are clearly dictated by genes, and it would be impossible for natural selection to not act on such dynamics, vis-a-vis co-evolution, that is the evolution of the genome, in tandem with the evolution of the environment.

We posit that free energy principle can explain many of the group-scale adaptive traits that have evolved in plant, animal societies, and ecosystems despite the lack of imagination some biologists suffer from. Many specimens cannot thrive in isolation, group organization reduces risks, facilitates resource sharing, intelligence amplification and division of labor. In herds, while some specimens observe the environment for predators, some will keep feeding. In human tribes, there is usually a sexual division of labor, but other kinds of labor division exists in many primate and non-primate species. Chimpanzees engage in cooperative hunting, Boesch et al. argue that this behavior is due to mutualism [3].

4 Symbiosis as a Naturalist Foundation of Ethics

Philosophers sometimes suffer from a lack of imagination, especially when they wish to present a general theory of ethics. Although consequentialism is more or less the only game in town for naturalist philosophers, simplistic theories such as utilitarianism have not resulted in a sensible explanation of ethical behavior. Utilitarianism of any sort is problematic however, the popular and simplistic positive/negative utility proposals of Singer and his followers that try to reduce utility to pain/pleasure are particularly bankrupt theories due to the utility monster objections – see a discussion of utility theory in [2].

The main philosophical problem with utilitarianism is that it tends to

 doi:10.20944/preprints201908.0318.v1

be either an underspecified theory, not being able to explain what utility we should choose, or ineffective and plain wrong when they do try to propose a specific kind of utility, like positive utilitarianism which must be classified as a bland manner of scientism.

The second kind of wrong consequentialism is the insistence of "AI Safety" researchers on human preferences. Some researchers seem to have gotten the idea that they could make behaviorist reinforcement learning agents safe, if only they could make them imitate human preferences through arcane methods such as inverse reinforcement learning that will be used to infer reward functions. Humans do infer the mental states of others, including their goals, and learn from their behavior, including ethical constraints, although there is no scientific evidence that we infer reward functions, or that reward functions correspond to ethical decision making. Human preferences have however been successfully integrated into a reinforcement learning model recently [5].

Nevertheless, there is a massive epistemological obstacle to any such method that might make utility maximization agents safe. There is nothing universal, stable, or well-defined about human preferences. Humans have many meaningless, vague, contradictory, irrational, harmful preferences. Moreover, the preference learning hypothesis does not extend to other species, neither does it seek to explain the commonality of ethical behavior in other species and our own species, with many divergent social codes. A recent experiment that collected human preferences about ethical dilemmas in the self-driving domain illustrates this epistemological problem perfectly [1]. The decision to spare a baby or a grandmother apparently varies according to demographics, which indicates that mimicking human ethical decisions superficially will likely not be a safe solution for AGI agents that interact with humans. We cannot deflate ethics by reducing it to human preferences. The reason philosophers of ethics exist is because human preferences are often not good or reliable. A compelling consequentialist account of ethical cognition is due to Dennett in his seminal book Freedom Evolves [7], however, even an evolutionary account might fall short of explaining the universal principles underlying ethics, which the free energy principle might allow us to accomplish.

Pain and pleasure are merely proprioception that the organism uses to detect generally favorable, and unfavorable conditions the organism finds itself in. They are guiding internally generated sensory signals, and they are only part of a heuristic evolved to either drive the organism towards such favorable states, or deter from the unfavorable states. However, unless you are a radical behaviorist, there is no reason to assume that is all there is to cognition and/or behavior. Behaviorism has been refuted strongly by Chom-

 doi:10.20944/preprints201908.0318.v1

sky, and we generally think that there is little merit left in the behaviorist school.

Animals use a variant of the perception-cognition-action cycle, and it is particularly in cognition that ethical behavior may emerge; it surely cannot be explained away as habitual behavior, learning from elders, imitating others, and so forth. Those explanations would be circular, not explaining why some behaviors are better than others, failing in another way than utilitarianism. Even more importantly, such an approach cannot explain why the human brain evolved to make contextual and ethical judgement possible.

We propose that the universal modes of co-existence, called symbiosis and parasitism are well-suited to construct a more empirically grounded and universal theory of ethics based on biological research rather than armchair philosophy. The classical examples of symbiotic behavior involves pairs of animals, which exchange benefits in a mutualist economy. However, the common description of symbiosis is perhaps not sufficiently broad. Impressive examples of symbiosis have resulted in the modern cell metabolism and multi-cellular organisms. Mitochondria and the cell, and cells within an organism have a symbiotic relationship. The hypothesis that mitochondria and chloroplasts descended from specialized bacteria is called symbiogenesis or the endosymbiotic theory [19]. In our view, symbiosis extends to social organization and even extra-somatic organization in species.

Once upon a dead world, it was through symbiosis that nascent forms of life flourished and then exploded in variation and population into the environment. Within the cybernetics approach of the free energy principle, then, the generalised symbiotic relationship could be considered as any feedback loop that reduces free energy over an ensemble of organisms. The symbiosis may be regarded simply as pooling resources, but it may also result in stable new forms like that of the mitochondria. Friston et al. explain that two organisms that engage in communicative, collaborative acts must seek to co-operate effectively to reduce free energy. Such symbiotic behavior explains why the main use of language is to transmit knowledge and imperatives, indeed it explains why language evolved at all [12]. If both organisms share their inferences about the world, they will quickly build a more accurate collaborative world model, and be able to plan better with exchanged information (world-model parts), and with imperatives, they can execute collaborative plans [12]. We submit that symbiosis does not stop at the boundary of two individuals, or even entire societies and populations. Symbiosis extends to ecosystems; the myceline networks form a mutual immune system for the trees in a forest. Symbiosis extends to the entire biosphere, the whole carbon cycle is an instance of symbiosis at a planetary scale.

Therefore, the phenomenon of altruism cannot be dismissed as a conse-

 doi:10.20944/preprints201908.0318.v1

quence of self-interest, as some contemporary naturalist thinkers like Dawkins [6] and Pinker [22] seem to believe – in their futile attempt to justify a dysfunctional economic ideology that benefits parasitic behavior. If that were the case, the members of a human society would not readily distinguish between cases of symbiosis and parasitism, however, they do, and they do in a dramatic manner, which is the fundamental debate that underlies much of political philosophy. While a worker co-operative is a symbiotic economic structure, the billionaire class collects rent, and extracts undue profit from every sector, which *might* be considered parasitic. Social welfare or altruism was not invented by enlightenment era thinkers; it existed well before humans evolved. Many species survived solely due to their particularly rich ways of exploiting social relationships. The quixotic attempt to reduce ethics to selfishness remains as a tacit approval of social darwinism and anti-social political ideologies. See Wilson's paradigm changing book on the new field of sociobiology for a contrastive view that elegantly explains social instincts such as attachment without succumbing to such basic fallacies [25].

Thus, symbiosis may serve as a suitable, naturalist foundation for understanding ethical problems from the ground up. Our species evolved *many* cognitive traits to behave kindly, affectionately, socially; these were not accidental features, had there been no need for symbiotic behavior, our species could easily take on a more solitary nature. Yet, we delight in the company of others, and have an insatiable need to care for our family members and those in plight – you may find a discussion of social cognition in [18]. These instincts stem from the simple fact that symbiotic behavior results in fitter populations. In other words, distributing resources fairly while dividing labor effectively has been a winning strategy for our yet young intelligent species. Slavery is indeed worse than a society that lets individual agents flourish yet benefit from a symbiotic co-existence, because the imbalance in benefits produces suboptimal outcomes, not because it is individually better for one person to avoid slavery; it matters if a great number of slaves are freed rather than one, as far as consequences go.

Moreover, the theory might allow us to realize generalized robot laws as we proposed in [21]. Once we have a universal definition of life, and the constraint which keeps organisms alive (minimizing free energy), it might be possible to formulate an AGI agent that can recognize people, and keep them away from harm. The most plausible meta-goal considered in [21] is also symbiotic: to preserve and pervade life, and culture throughout the universe. It might be the case that such cosmic symbiosis meta-goals follow from the free energy principle.

 doi:10.20944/preprints201908.0318.v1

5 Equanimity as the Final Goal of Organism

There are rare moments of satisfaction in one's life, when seemingly every need has been met. Cognitively, such satisfaction for an active inference agent is observing a state of minimum free energy, which entails that the agent has sufficiently complete information about the environment to predict future states, and that it can comfortably manipulate about the environment to produce likewise minimum free energy states. The trajectory of the environment is still burdened with Brownian motion, but there are events where it seems that some manner of dynamic equilibrium has been reached within a microcosm. Such moments may be termed peaceful, and coincides with the intuition of *equanimity*. The more an agent knows about future states, the less anxiety it will experience. Therefore, in the most desirable states, the organism will engage harmoniously with the environment, and be capable of withstanding surprise input in the future.

When one is situated in a relatively calm natural environment with a few well-controlled stochastic processes such as a gently flowing stream of water, the mind can experience the satisfaction that accompanies a complex information source that is nonetheless modeled accurately by the agent's world model, and where little surprise is expected. Hence, the Zen garden acts as a model of biology's final goal. The active inference agent will cultivate the environment until it attains a sufficiently high degree of free energy minimization, which explains the appeal of introvert behavior and nature retreats. Such introversion is however balanced out by the innate drive of curiosity that urges the agent to continually acquire new information about the environment; after a while even the most exquisite and calming Zen garden will tend to boredom for the agent [13].

Another fundamental prediction of the free energy principle is that despite the noise in the environment, the trajectory of the agent will tend to carve out valleys of fitness and gradually ascend to higher plateau. This observation explains why the organisms also seem to have a desire to improve their conditions, and vary and multiply, instead of being satisfied with them by default.

As such, the agent must both thrive and participate in ever greater feedback loops as well as experience harmonious, satisfied labor. The variety of sub-goals implied by the universal goal of free energy minimisation paints a more familiar picture than alternative agent theories. Thus, the free energy principle gives way to a primary directive of life, and intelligence. It is conceivable that there are other significant aspects of life, however, only one is universally fixed.

 doi:10.20944/preprints201908.0318.v1

6 Towards the World-Mind

The principle of free energy does not stop at arbitrary boundaries and scales, therefore, we may hypothesize that a future world-mind, connected through cybernetic and artificial intelligence technology, will be more effective at reducing free energy than independent units. The advent of the internet may be regarded as a first step towards such global cognition. Even if the current world-mind has no mental features, it still operates as a weakly coupled collaborative system of sorts through arrangements of communication and economics. On the contrary, this global human mind is apparently lacking enough cohesion, else it would act swiftly to deal with anthropogenic global warming. Therefore, a world-mind through human units would likely not be effective enough for the cognitive requirements of a world-mind.

Artificial units with much better energy efficiency (which we expect to appear after 2030 due to Koomey's law [16]) are more likely to result in an effective world mind. The energy efficiency of computation, as well as the capabilities of robotic technology will be critical in the free energy reduction of such an organism, therefore its genesis is a step towards the hypothetical infinity point (singularity), whereby positive feedback loops started by AI drive improvements in AI technology itself towards physical limits of computation. Even if some ultimate computer or AI close to the physical limits of computation [20] is not achieved within a short timespan as predicted by Ray Solomonoff's infinity point theory [23], a more reliable outcome could be the emergence of global-scale autonomous AI organisms.

Speculatively speaking, the free energy principle would compel such a world-mind to assimilate biological information of all sorts to improve its fitness, therefore we might find that technological immortality via brain simulation may emerge as the natural path before such an organism evolves. The simulated human brain would find it much easier to develop high-bandwidth cross-connections and coalesce into a heterarchical super-brain by modifying their neural architectures. Evolution in silico would progress millions of times faster than biological evolution, and superior post-human organisms would evolve rapidly at any rate. In other words, the Internet would finally come alive and develop awareness. We cannot yet completely fathom what an organism with a billion minds and bodies would be like, however, we can estimate that it might be better at survival than we are. Such an organism would then find it imperative to migrate to other worlds to prolong chances of survival, and therefore the interstellar expansion of a society of world-minds would have begun.

There is a speculative idea called "Gaia hypothesis" in biology that the Earth itself may be considered a biological entity. According to the free

 doi:10.20944/preprints201908.0318.v1

energy principle, that is not quite true, as Earth is more like a cosmic womb, or a seed pod, which brews life, although the planet does have an atmospheric boundary, and many synergistic processes, which perhaps makes our pale blue dot more like a proto-cell. Earth has the potential to be a world-mind, though it is not yet one. The planet barely has any sensory states or can influence other planetary bodies in any significant way, it is mostly inert in terms of animacy, and unfortunately the highest intelligence density is that of humans on it. Perhaps, one day, it shall saturate with intelligence, and it shall spawn the seeds of a cosmic dandelion to disseminate life to the rest of the Cosmos.

7 Discussion

We reviewed the free energy principle as a new scientific definition of life and autonomous cognition in terms of its epistemological and ethical implications. Epistemologically, the approach supports a flawless reductionism that reduces life and intelligence to the lowest possible level which is physical motion. Unlike previous scientific definitions of life, it is not peculiar to Earth biology (in terms of DNA, reproduction, etc.), but it is a physical definition that is applicable to any adaptive system. The definition does not fail at systems close to the boundary between animate and inanimate such as adapting crystals, virii, and cryostatic organisms.

Remarkably, the free energy principle addresses the biogenesis problem by introducing a completely universal theory of self-organization. In this view, the universe is tilted slightly in favor of life and intelligence, rather than shooting darts at a cosmic board. Such conclusion urges us to revise our view of extraterrestrial intelligence, which is mostly couched in earth centric concepts such as our search for exoplanets that are similar to Earth. Under the free energy interpretation, the Cosmos can and will produce fundamentally different forms of life, given sufficient resources. Some of these forms will inevitably be carbon based, but others might be based on silicon, or even nuclear reactions, or other physical interactions.

The ethics implied by the free energy principle belongs to consequentialism, as the active inference agent continually projects the future, and plans to minimize free energy. The universal goal of the agent is survival, however, the goal naturally extends to groups of organisms, and the fitness of a group may be evaluated in its capacity to collectively minimize the free energy of the population. Individually, the free energy principle manifests as homeostasis, curiosity, creativity, habit formation, and exploration, while socially it manifests as symbiosis, attachment, communication, empathy, altruism,

and ethics, among other phenomena mercifully forgotten, for the entire list would be too long to include. Those populations with better fitness will be selected over others in the long run. We argue that the free energy view naturally gives rise to a preference of symbiosis over parasitism, as it is a general strategy that improves the fitness of groups. Just as an organism riddled with parasites is not likely to survive long, a society riddled with parasitic behavior will not reach optimal lifespan or fitness. We also argue that the subjective state of mind entailed by the principle is that of *equanimity*; peaceful and harmonious co-existence with the environment.

Finally, we argue that another natural inclination implied by the principle is rapid evolution towards a world-mind incorporating biological and artificial elements. We argue that such a world-mind would necessarily try to incorporate as much biological information as possible in order to minimize its expected free energy, optimize its technological base to maximise energy efficiency, and additionally choose collective, efficient cognitive processes to guide future actions, which will at a first approximation look like direct democracy and brain simulations co-operating over the Internet. Therefore, the free energy principle seems compatible with the infinity point hypothesis, also known as the singularity.

References

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[2] Salvador Barbera, Peter Hammond, and Christian Seidl. *Handbook of Utility Theory Volume 1: Principles*. Kluwer Academic Publishers, 1998.

[3] Christophe Boesch, Hedwige Boesch, and Linda Vigilant. *Cooperative hunting in chimpanzees: kinship or mutualism?*, pages 139–150. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[4] Noam Chomsky. A review of b. f. skinner's verbal behavior. *Language*, 35(1):26–58, 1959.

[5] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv e-prints*, page arXiv:1706.03741, Jun 2017.

[6] Richard Dawkins. In defence of selfish genes. *Philosophy*, 56(218):556–573, 1981.

[7] Daniel Clement Dennett. *Freedom Evolves*. Viking Press, 2003.

[8] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 2010.

[9] Karl Friston. A free energy principle for biological system. *Entropy*, 14(11):2100–2121, 2012.

[10] Karl Friston. Life as we know it. *Journal of The Royal Society Interface*, 10(86), 2013.

[11] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John ODoherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience and Biobehavioral Reviews*, 68:862 – 879, 2016.

[12] Karl Friston and Christopher Frith. A duet for one. *Consciousness and Cognition*, 36:390 – 405, 2015.

[13] Karl Friston, Christopher Thornton, and Andy Clark. Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3(130), 2012.

[14] Marcus Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel and C. Pennachin, editors, *Artificial General Intelligence*, Cognitive Technologies, pages 227–290. Springer, Berlin, 2007.

[15] Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, 15(138):20170792, 2018.

[16] Jonathan Koomey, Stephen Berard, Marla Sanchez, and Henry Wong. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, 33(3):46–54, 2010.

[17] Tor Lattimore and Marcus Hutter. No free lunch versus occam's razor in supervised learning. In David L. Dowe, editor, *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence - Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, VIC, Australia, November 30 - December 2, 2011*, volume 7070 of *Lecture Notes in Computer Science*, pages 223–235. Springer, 2011.

[18] Matthew D. Lieberman. *Social: Why Our Brains Are Wired to Connect*. Crown, 2013.

 doi:10.20944/preprints201908.0318.v1

[19] Lynn Margulis. Symbiogenesis and symbionticism. *Symbiosis as a source of evolutionary innovation*, pages 1–14, 1991.

[20] Eray Özkural. Ultimate intelligence part I: physical completeness and objectivity of induction. In *Artificial General Intelligence - 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings*, pages 131–141, 2015.

[21] Eray Özkural. *Philosophy of Mind: Contemporary Perspectives*, chapter Godseed: Benevolent or Malevolent? Cambridge Scholars, 2017.

[22] Steven Pinker. The false allure of group selection. *The handbook of evolutionary psychology*, pages 1–14, 2015.

[23] Ray J. Solomonoff. The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, 5:149–153, 1985.

[24] Peter Sunehag and Marcus Hutter. Intelligence as inference or forcing occam on the world. In Ben Goertzel, Laurent Orseau, and Javier Snaider, editors, *Artificial General Intelligence - 7th International Conference, AGI 2014, Quebec City, QC, Canada, August 1-4, 2014. Proceedings*, volume 8598 of *Lecture Notes in Computer Science*, pages 186–195. Springer, 2014.

[25] Edward O. Wilson. *Sociobiology: The New Synthesis, Twenty-Fifth Anniversary Edition*. Belknap Press, 2000.