Diverse intelligences: a research proposal

Asvin G

September 2, 2023

1 Introduction

I have recently become extremely interested in understanding the fundamental nature of intelligence and the capability/possibilities for intelligent agents to co-operate and form a distributed sense of self/collective goals (motivated by GPT-4 and a sense of where such technologies might take us). In thinking about this problem and searching the literature, I have found Dr. Michael Levin's work to be the most insightful and foundational for thinking about these topics. I find the following definition of his very useful as a starting point.

Definition. *Intelligence is the ability of a system to attain its goals despite interventions and obstacles. Implicit is the notion of a problem space in which the system is active, which could be physical space, "mathematical space", morphological space and so on.*

One of the key questions I would like to explore is what is *necessary* for a system to behave in an intelligent manner. Could we say anything about the system's internal workings given that it appears intelligent (in a given problem space). Alternatively, does this place any constraints on the problem spaces? As I will discuss in §2 (and as Levin and others point out in various places), the notion of computation seems to be crucial here. Moreover, it gets at fundamental problems about what computation really is that remain open still.

The second aspect of Levin's work that I am extremely excited by are his various experiments on the mechanisms by which cells (individual agents in their own right) manage to form a unified sense of self and co-operate in the service of a "top-down intelligence". At my most ambitious, I would like to develop a theory of "ethics" for how ecosystems of intelligent agents should behave, i.e., how should the system be engineered in order to maximize co-operation and goal alignment in this ecosystem.

2 Intelligence as compression

In this section, I want to explore what claims, if any, can be made in general about intelligent agents. As a beginning point, we will use Definition 1. More precisely, by *problem space*, we will mean simply any set X (or set with structure such as a manifold equipped with a fitness function) and by an agent, we mean a dynamical system $T : X \to X$. In other words, points on X encode the "state" of the agent and the evolution function T encodes the dynamics of the agent. Moreover, one might invoke an "energy" functional $E : X \to \mathbb{R}$ and the "goal" of the agent is to minimize E (which might correspond to maximizing fitness).

An intelligent agent is then an agent which prefers certain states $S \subset X$ and has the ability to navigate to these states S from a "large" fraction $F \subset X$ of the problem space. Examples of intelligent agents include at a high level, "humans" navigating societal spaces or physical spaces in order to achieve their goals, artificial intelligence systems like alpha-zero learning to play go or chess but also at a much lower level, cells navigating morphological space in order to build and maintain limbs or evolution navigating the space of genotypes.

Remark. *This formalism is only meant to serve as a very rough starting point given its many deficits. Most importantly, this model does not encompass an*

ecosystem of interacting intelligent agents which is our eventual goal. Perhaps something like David Spivak's formalism will be useful here.

In the rest of this section, I would like to explore the following hypothesis:

Hypothesis. *Any intelligent agent has to have some internal representation $\mathcal{R}$ of the problem space X based on which it can compute the "optimal" updating function T.*

This is only a preliminary hypothesis and therefore, I would like to be appropriately vague about the various terms "internal representation" and "compute". Nevertheless, there is strong evidence that something like this must be true. For instance, one can prove theorems [3] showing that if an agent X wants to control the dynamics of another system Y and do so well (in a precise information theoretic sense), then X must contain a *good* model of Y.

Due to physical constraints, it is often not possible to have a perfect representation of X and $\mathcal{R}$ necessarily has to lose information and moreover, computational resources have to be invested in the process of decryption or decoding the model's outputs in real world actions.

This hypothesis stands in contrast to a long philosophical tradition (especially in the west) where "thinking" is taken to be a perfectly rational, transparent process governed by logical rules and perhaps some heuristics. On the contrary, I think the latest advances in machine learning suggest a contrary view: intelligence is often a complete blackbox, even to humans, and consists of constructing an internal computational model which *produces* good responses to input stimuli (or hypothesized scenarios). In light of this, I think the following conjecture deserves serious investigation.

Conjecture. *Compression is all you need for intelligence.*

Unpacking the above conjecture a little, I think it is plausible that any *sufficiently* good compression of *sufficiently* interesting data will naturally have good generalization properties. The motivating example are GPT style LLMs. In a very concrete sense, GPT-4 is a good compression of its training corpus - it recovers this corpus with high fidelity when prompted and this was the only metric on which GPT was trained. Surprisingly, GPT has excellent generalization qualities to all sorts of unexpected domains such as coding, general pattern recognition [5] etc.

The conjecture suggests two main lines of enquiry. First, it is clear that the problem domain that we are operating in is of paramount interest. Trying to compress random noise will surely not lead to anything interesting! Similarly, theoretical computer science suggests that good encryption is almost certainly possible and such encrypted data should in principle, make learning patterns in the data computationally infaesible.

Question 1. *Given these facts, what about "real world data" makes it so amenable to compression (and learning)?*

Some recent research in the machine learning literature suggests some necessary properties. For instance, GPT type models exhibit few-shot learning, i.e., learning new concepts from a few examples without changing their internal weights. This recent paper [2] investigates whether certain kinds of training data help learn this skill. They find that indeed, few-shot learning emerges when the training data has "clumps" or "burstiness" where similar concepts occur together. Similarly, it also helps to have large numbers of rarely occuring features that need to learned (in the training set). Can we find analogues of this in biological agents?

More speculatively, it seems that real world data is computationally compressible because it is *generated* by computational processes in the first place. It seems to me to be a deep question why regularity is so favored in real life, building on Wigner's "The unreasonable effectiveness of mathematics in the Natural Sciences".

Second, how can we tell that a compression is *sufficiently* good and how can a system learn such a good compression?

Question 2. *In general, it seems that both evolution [1] and the stochastic gradient descent we use to*

train LLMs [5] "like" to find general pattern matching machines that can then build internal models of the data they encounter rather than "hardcoding" a model that fits the training data. Is it possible to test this hypothesis directly in biological or AI systems?

Similarly, what possible strategies exist in building intelligence?

Question 3. *What search algorithms are possible in naturally occuring problem spaces? It appears that locality is a strong constraint - is the only option (stochastic) gradient descent?*

The cognitive science literature in particular seems to be prime mining ground here. The question of what exactly *mental representations* are seems to have much in common with the question under consideration here. For instance, [6] develops "*a theory of the underpinnings of symbolic cognition that shows how sub-symbolic dynamics may give rise to higher-level cognitive representations of structures, systems of knowledge, and algorithmic processes*" in an explicit enough manner to be computationally implementable. In other words, this paper attempts to build a learning system ("a language of thought") that can learn arbitrary patters.

Question 4. *Can we identify how a "language of thought" is implemented in biological organisms?*

2.1 Observer dependence

Taking computation as the central key to intelligence leads to several questions about the nature of computation itself. While there is a formal definition of a computation as something *done by a Turing Machine*, this is not completely satisfactory for an application to biology because of the varied ways in which computation can be *embodied*.

The classical definition of a Turing machine takes for granted that the various parts of it, such as the memory tape or the read head, can be clearly identified but this presupposes an intelligent observer. It is not always so clear what counts as a computation - for instance, should one count a falling rock as a computation of the laws of gravity? Given these difficulties, it is tempting to bite the bullet and try to build an observer dependant theory of biology.

There are also several suggestions from physics that an observer dependant point of view is justified. This is of course a central question in Quantum Mechanics but the problem also shows it's head in the more classical (and fundamental) statistical mechanics theory of entropy. The computation of entropy depends on how exactly one chooses to coarse-grain the object of study. Indeed, if one had *perfect information*, then there would be no change in entropy over time. A tantalizing possibility is to view entropy itself as a consequence of agents with bounded computational power observing events in a lossy way so that entropy corresponds to an *encryption* of the information one begins with.

3 Self Organization

I am interested in the broad question of how self-organization among intelligent agents is possible and moreover, how a coherent sense of self is generated and maintained over time. For instance,

Question 5. *Is the key factor in generating a group identity the* speed of information transfer *among the individual agents? Or does it have more to do with the "software"?*

Levin proposes sharing *stress* as a key mechanism in achieving co-operation. It would be extremely interesting to work out the mechanics of this. After all, when a virus induces stress in its host cells, the host responds by identifying the virus as an outsider and trying to eradicate it which is the polar opposite of co-operation. On the other hand, some "parasites" do manage to cohabit in a host organism (such as our gut bacteria). This suggests another question:

Question 6. *Is the mechanism of co-operation different between cells that share the same genetic code vs those that don't? In other words, is there a significant advantage to co-operation when beginning with the same "hardware"?*

One lesson of Levin's body of work is that once one is open to the existence of unconventional intelligences in a diverse range of problem domains, one simply finds it everywhere. Levin has exploited this to great effect in the biological realm and I am interested in transporting these insights to different realms (and vice-versa).

For instance, one perspective on the global economy is as a distributed computation for finding the *values* of various goods. This process is truly non-local in that no individual human or even small groups of human have much influence over the value of any good relative to any other and yet, prices do get decided at both a local and global scale. This process is even quite efficient and opportunities for arbitrage are relatively rare. *Can one find analogs of this in biological networks?* And can we otherwise use insights from economics to help understand the questions outlined above?

Question 7. *How does the agent store its goals? Can the goals be arbitrary subsets $S \subset X$ or are they constrained to be "local minima" for natural energy functionals E?*

4 Miscellaneous questions

In this section, I simply collect some questions that stuck me while reading through Levin's research.

Question 8. *Since life is agential and behaves intelligently at every scale what, if anything, is special about human intelligence? Is a matter of having extremely adaptable goals?*

Question 9. *Can xenobots be made from skin cells with differing genomes?*

Question 10. *Can we identify the "competency genes" in planaria?*

Question 11. *What is the mechanism of memory storage in planria? Surely it must be stored in a highly redundant way in order to allow regeneration from any body part. Is memory always stored in this redundant way?*

There is a general feature of intelligent networks where "most" of the intelligence is localized in a relatively small part of the network. For instance, there is by now a large body of literature ([4], [7]) on *pruning* neural networks in order to make them sparse while retaining most of the capability.

Question 12. *Do biological networks show a similar behaviour? One hypothesis for why this might be the case that a large network helps avoid local minima (by turning them into saddles) during the model building process. Can this hypothesis be tested in biological networks?*

Question 13. *Since the "competency" of an organism can often account for genetic deficiencies, does this explain why organisms are able to obtain mutations that involve several genetic changes at the same time (where partial changes might be very maladaptive)? How does this "competency" interact with the standard evolutionary story of genetic fitness more generally?*

References

[1] Surama Biswas et al. "Gene regulatory networks exhibit several kinds of memory: Quantification of memory in biological and random transcriptional networks". In: *Iscience* 24.3 (2021).

[2] Stephanie Chan et al. "Data distributional properties drive emergent in-context learning in transformers". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 18878–18891.

[3] Roger C Conant and W Ross Ashby. "Every good regulator of a system must be a model of that system". In: *International journal of systems science* 1.2 (1970), pp. 89–97.

[4] Elias Frantar and Dan Alistarh. "SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot". In: (2023).

[5] Suvir Mirchandani et al. "Large Language Models as General Pattern Machines". In: *arXiv preprint arXiv:2307.04721* (2023).

[6] Steven T Piantadosi. "The computational origin of representation". In: *Minds and machines* 31 (2021), pp. 1–58.

[7] Mingjie Sun et al. "A Simple and Effective Pruning Approach for Large Language Models". In: *arXiv preprint arXiv:2306.11695* (2023).