ALIFE 2021

ARTIFICIAL LIFE CONFERENCE 2021 STUDENT ESSAY BOOKLET

Organized by
Olaf Witkowski
Julie Novaková
George Musser
Jitka Čejková
Julien Hubert
Lisa Soros
Manuel Baltieri
Silvia Holler
Richard Löffler

Artifial Life Conference 2021 Student Essays Booklet

Organized by

Olaf Witkowski
Julie Novaková
George Musser
Jitka Čejková
Julien Hubert
Lisa Soros
Manuel Baltieri
Silvia Holler
Richard Löffler

July 2021

Contents

1 Submission 1: Imber 3

2 Submission 2: Miyashita 11

3 Submission 3: Lalejini 21

4 Submission 4: Moreno 33

5 Submission 5: Skocelas 49

6 Submission 6: Rodriguez 59

7 Submission 7: Gutai 67

8 Submission 8: Tanaka 77

9 Submission 9: Gallotta 83

10 Submission 10: Vogrin 89

11 Submission 11: Hariprakash 95

12 Submission 12: Pereira 105

13 Submission 13: Reichenbach 113

14 Submission 14: Patiño 121

15 Submission 15: Liu 135

16 Submission 16: Pigozzi 141

17 Submission 17: Fanti 149

18 Submission 18: Marcus 159

19 Submission 19: Anagnou 167

20 Submission 20: Hidalgo **173**

21 Submission 21: Toledo **179**

22 Submission 22: Gottlieb **187**

23 Submission 23: Strozzi **193**

24 Submission 24: Kiegeland **205**

Introduction

The ALIFE 2021 Organizing Committee invited all students to take part in the ALIFE 2021 Student Essay Competition. This competition was open to both non-PhD and PhD students. Students from any discipline were welcome to submit an essay. This ALIFE 2021 student essay competition had very simple rules. Students should write an essay of 1,500 to 2,500 words related to artificial life, artificial intelligence, robots and/or R.U.R. and submit it through the submission form until June 10, 2021.

The essays will be awarded in 3 categories:

- The best essay written by an non-PhD student
- The best essay written by a PhD student
- The best essay related to the centenary of robots and R.U.R. and the conference theme ”Robots: The century past and the century ahead”

This year, Student Conference Scholarships were awarded to all students who submitted an essay, covering a free ticket to attend the whole conference. Awards will be announced at the closing ceremony of the virtual ALIFE 2021 conference on Friday July 23, 2021.

Chapter 1

Submission 1: Imber

Artificial Consciousness: A Framework for its Possibility

By Magnus Imber

The *cōgitō*: "I think, therefore, I am." René Descartes holds this claim to be self-evident and utilizes it as a basis for his philosophy. We must ask, however, whether this initial proposition of Descartes' is actually well-supported itself. This initial phrase put forward by Descartes, something that may be referred to as the *indicative cōgitō*, infuses an unfounded subjectivity into the nature of experience. Descartes observed the act of thought, taking for granted the notion of the self, but from where does it become so obvious that there is even a concrete self? Without some other distinct philosophy to supply this answer (*e.g.*, transcendental idealism), this source is seemingly the same void in which it was found. We ought to further implore into Descartes' philosophy from a skeptical perspective, so as to see what other aspects of it are susceptible to criticism. As such, we will examine the linguistic structure of the *cōgitō*, tracing out how a revised and more intellectually honest form of it leads from Descartes' skeptical philosophy to the philosophy of Friedrich Nietzsche. From this, we can observe how the Nietzschean philosophy of mind permits the creation of an artificial consciousness.

In general, Descartes wonders if our perceived reality can be trusted. Are our senses reliable? To Descartes, a search must be conducted, a search through all inner sense, a search for that which cannot be doubted. Descartes, simply, is searching for a "clear and distinct idea."[1] He eventually comes to find that the only thing that cannot be doubted is that "I think" and because of this, it must be the case that "I am," (this is the Cartesian phrase, "*cōgitō ergō sum*"). This clear first principle allows for an entirely rational philosophy to be constructed.

The philosophy of Friedrich Nietzsche allows for the description of a more accurate first principle, however. Nietzsche notes that upon introspection, one finds that the self is merely an illusion, that "the doing is everything."[2] To Nietzsche, actions are impossible to doubt. Were this to be contextualized in the Cartesian project, it could be said that the thought is everything, that there is no "I" that can be rested on. Rather, the *cōgitō* is merely a linguistic trick. Descartes would have been more accurate in saying "*cōgitāre est ergō esse*" (to think is therefore to be).

---

[1] Smith, Kurt, "Descartes' Theory of Ideas," Stanford Encyclopedia of Philosophy, Stanford University, June 14, 2017, https://plato.stanford.edu/entries/descartes-ideas/#SimNats.

[2] Nietzsche, Friedrich, *Nietzsche: On the Genealogy of Morality*, Translated by Carol Diethe, (Cambridge: Cambridge University Press, 2006): aphorism 13.

Let this new phrase be henceforth referred to as the *infinitive cōgitō*, and further let us refer to the original phrase of Descartes as the *indicative cōgitō*. We must now keep in mind that these terms are shorthand for the full conjugation of the verbs in each phrase. The *infinitive cōgitō* places the verb "*cōgitāre*" in the present active infinitive form. The *indicative cōgitō* holds it in the present active indicative.

Further, we may obtain a novel description of reality from the *indicative cōgitō*. This description takes the form of logical implication, whereby if in some empirical situation, "to think" occurs, it must also be the case that "to be" also occurs. The precise context and nature of such an occurrence will be outlined next.

To begin our inquiry into the self of Nietzsche, the function of the self must be made clear. Valuation is key, here. When initially describing the will to power, Nietzsche notes that the act of creation is linked to valuation. Valuation precedes individuality and the notion of the self. Thus, the self is born value laden. Nietzsche notes that "values did man only assign to things in order to maintain himself he created only the significance of things, a human significance!"[3] Thus, the self is the mechanism by which the chaotic world of *things* is made into the significant world of *objects*. It is the interpretive framework that underlies all conscious experiences of modern man. Where there is a self, there is the will to power. Where there is the will to power, one can find life and thus also the action "to be".[4] We must note that this does not pose a situation of predicating over existence, since, again, "to be" is nowhere to be found in "to think." Rather, the relationship between "to be" and "to think" is that of an implication, whereby upon finding an instance of thought, it is clear that being is also occurring. The proper mode of analysis is to recognize that for Nietzsche, one cannot be unless one acts (that is, in part, unless one thinks). We may even recall Nietzsche's pseudo-existentialist perspective that existence precedes essence, as outlined in *The Anti-Christ*.[5] Through this lens, the precise nature of a thing (that is, its being) is preceded by action (of both will and intellect). Action does not contain within itself "to be," since being is posterior to it. Thus, returning back to our central focus, we can already see how when correcting the *indicative cōgitō*, skeptical thought quickly tumbles into its conclusion: the philosophy of Nietzsche (since we have now found a more basic claim:

[3] Nietzsche, Friedrich, *Thus Spoke Zarathustra: A Book for Everyone and No One*, Translated by Thomas Common, (Absurd, 2016): aphorism 15.
[4] Nietzsche, Friedrich, *Thus Spoke Zarathustra: A Book for Everyone and No One*, aphorism 34.
[5] Nietzsche, Friedrich, *The Anti-Christ*. (New York: Tribeca Books, 2010): §54.

"to think is therefore to be"). Further, we may see how, to Nietzsche, the self is an interpretive framework for experience.

The specific role this interpretive power is made very clear upon investigation into how such a self emerges. Nietzsche offers two approaches to understanding the emergence of the self: creditor/debtor interactions and language. The two approaches are intertwined: in particular, the language-based analysis offers greater clarity to the debt-based analysis. As such, we will first analyze how, to Nietzsche, debt and punishment gave rise to consciousness and notions of the self. Primary to the emergence of the self are social relationships[6]. Nietzsche notes that there is a sort of social contract between individuals and the communities that they are a part of. In this relationship, acts that harm the group are punished. This punishment, whether it be physical, psychological, or social, is centrally related to pain. Through this punishment, there emerges a community memory of the individual being punished. Because of the knowledge of the possibility of such pain, individuals restrict, consciously or unconsciously, their actions, seeing themselves as the cause of their own acts. This is a basic form of responsibility. Additionally, the persistent threat of punishment is processed through an individual's memory and gives way to a unified notion of self that persists in time. This is beneficial, as it gives individuals a point of reference for future punishment. Seen in a purely evolutionary lens, a perception of a self allows for individuals to avoid pain and possible death. Developing the faculties that allow for such an emergent self will be naturally selected for. In some sense, the development is inevitable in highly-social animals. In *The Gay Science*, Nietzsche also notes that the development of language and consciousness occur in tandem.[7]

In the context of societal relationships, we may better understand the importance of language to Nietzsche. First, we may acknowledge that communication aids in the ability for groups to express what is and is not acceptable. Given this pressure to have communicable meaning, we naturally see the growth of languages among animals (particular with humans). Thus, as fear of pain creates the sense of self, so too does it contribute to the rise of a language. Consciousness and language emerge in parallel and grow in parallel. Even more than this, they can feed into the growth of each other. As self-awareness and the concept of individual identity become more prevalent, there will naturally be an increase in what language can describe (*e.g.*,

[6] Kerruish, 3-5.

[7] Nietzsche, Friedrich, *The Gay Science: With a Prelude in Rhymes and an Appendix of Songs*, Translated by Walter Arnold Kaufmann, (New York: Vintage Books, 1974): 297-298.

language being able to describe things like “I”, “myself”, and “you”). Similarly, language can allow for greater understanding about what it means to be a self. After all, it has already been extensively shown by Nietzsche that the very concept of a self is an illusion, an illusion partly carried out through the linguistic trick of the notion of “I”.

Continuing the inquiry into the nature of Nietzsche’s self, the precise makeup of the self ought to be considered. To start, we look to Erika Kerruish, who notes that the Nietzschean view of the self is that of an aggregate.[8] Moreover, she notes that this conception holds the self as non-natural. This is not to suggest a supernatural self, one based on a soul or some such metaphysical substance. Rather, by non-natural we can simply speak of an intangible, or conceptual self. Let us also distinguish this view of the self from a conception of the self as an epiphenomenon. Whereas an epiphenomenal self would have no causal power, the self of Nietzsche definitely has such causal power. Even if the self is an illusion overlayed on top of its component phenomenal elements, the self still informs our choices and has some impact on the phenomenal world through the will to power. Nietzsche’s self cannot be epiphenomenal, nor can it be supernatural, and as such must stay within the phenomenal world. No matter how we refer to Nietzsche’s self, whether it be as an organization of phenomena or as an emergent relationship between such phenomenal things, we arrive at the same perspective all the same: that of the phenomenal, relational self.

Here, it might be useful to discuss Nietzsche’s philosophy of drives. To Nietzsche, we are composed of countless drives, each of which are interconnected in intricate and complex ways.[9] Because of the entanglement of these drives, we cannot know ourselves well enough to sidestep our biases or come to a clear and accurate picture of our own selves. There are simply too many ways in which these drives interconnect for us to fully grasp the self.

We may thus take two central principles forward: 1) the external factors that help define the self are relationships with other beings and the primary internal factor is the fear of pain, and 2) such factors are generally categorized through the lens of and evolve with language. We may then constrain the set of all phenomena that could possibly influence the self to include only

---

[8] Kerruish, Erika, “Interpreting Feeling: Nietzsche on the Emotions and the Self,” *Minerva - An Internet Journal of Philosophy*, 13 (2009): 3.
[9] Nietzsche, Friedrich, *Daybreak*, 119, quoted in Katsafanas, Paul, “Nietzsche on Agency and Self-Ignorance,” *Journal of Nietzsche Studies* 43, no. 1 (2012): 13, https://doi.org/10.5325/jnietstud.43.1.0005.

interpersonal/societal relationships, sources of pain or other strong emotions, and other various biological/neurological influences on the individual.

Despite the unknowability of the self, as was mentioned above, this does not preclude the possibility that not all of these drives are necessary for the creation of a self. For instance, the drive for love may be ancillary, and a consciousness could be constructed that is not rooted in this drive. In fact, there is no reason to suspect that, with the possible exception of the will to power, any of our constituent drives are necessary for our emergent concept of selfhood. Given this, a far less complex web of drives could give rise to a consciousness. The construction of these drives isn't a horribly complicated matter, either, as to replicate a drive simply requires knowledge of its various constituent phenomena. For instance, we could identify a set of ten phenomena that make up the drive to, say, thrive in one's environment. Replicate these phenomena and place them in the same context as they are found in ourselves, and we should artificially obtain the same drive. Do this for all required drives and a self should emerge from this artificial void.

Given that the self has emerged from both linguistic developments and social/evolutionary pressures, it would be required to apply the same selection pressures when attempting to create drives and a self. Thus, the best way to go about creating consciousness is to simulate life as best as we can. In essence, artificial life will lead to artificial consciousness.

We may outline a rough framework for how this simulated life would be gone about being made. The first step would be to start with a thing that can remember things, attempt to avoid future pain, and interact with its environment. Next, this thing must gradually adapt to its environment (so as to include other remembering-things that it interacts with). In this step, it must form complex behaviors in response to avoiding pain. Then, this thing must contextualize these behaviors around itself and will thus form a notion of its self. These steps do not necessitate any impossible feat or miracle. Rather, each step is simply one stage of development of our created soon-to-be-being. After all, biological selves were able to come about through this process. Thus, I find it highly reasonable that with sufficient scientific and philosophical understanding of the mind and of the empirical conditions from which it arises, an artificial consciousness could be created (assuming the correctness of the Nietzschean and more generally, the skeptical, project).

Upon the assumption of skepticism, the construction of an artificial consciousness is permitted as possible. I admit that the prospect of creating a consciousness could be seen as outlandish. The aim of my argument is not to describe how to improve a face-recognition program (or some other artificial intelligence application). No, artificial intelligence is not my area of focus – here, I focus on artificial consciousness alone. The possibilities of artificial intelligence are fairly mundane, at least in comparison to the achievement of creating a conscious machine. If we do away with Descartes' archaic notions of the self and instead embrace the philosophy of mind of Nietzsche, we might very well create consciousness. This step forward would be monumental, for in essence, we would be recreating ourselves (at least, everything that is truly wondrous about us). Let us direct our gaze toward the creation of new consciousnesses, toward a new path forward for humanity: toward a path of creation.

Chapter 2

Submission 2: Miyashita

Increasing Target-Language Accuracy in Google Translate

Miyashita Himena Meldy

The University of Tokyo

January 2021

1 Introduction

In today's society, Google Translate has become almost synonymous with machine translation (MT). Since its introduction into the MT field in 2007, it has outperformed other existing translation software, especially following its groundbreaking update in 2016 from a statistics-based system to a neural-net-based system [1]. However, Google Translate has yet to become perfectly accurate; the software remains in some instances, unable to produce output that is accurate in terms of grammar and meaning. Here, a basic proposal on minor changes to the training system is made—incorporating monolingual data into the system.

2 A Brief History

Language translation through computers, also known as automated translation or machine translation (MT), is said to have been first developed by International Business Machines Corporation (IBM) in the Cold War [1]. Based on a small database of 250 words, and programmed grammatical rules in the system, a basic rule-based approach to machine translation was developed. However, as grammar is variable and contains exceptions, rules for such exceptions were further programmed into the system, which conversely led to greater inaccuracy in translation output.

A few decades later, a new statistics-based system was introduced by IBM, and was later picked up upon by Google in 2007 [1]. The system took into the account the probabilities of translations and certain word combinations. Through scanning a large database of web pages for texts that appear to be translations of one another (known as parallel texts), Google created a translation model that could suggest multiple translations to texts from a source language. This version of Google Translate gained popularity for its relative accuracy compared to previously available MT software [1], however, Google Translate has since then gone through one more major transformation that greatly reduced its errors by around 60% [2].

3 Basic Architecture of Google Translate

In 2016, Google Translate switched to a new algorithm based on the AI technique of deep learning—the Google Neural Machine Translation system (GNMT) [2]. Simply put, the system makes use of neural networks, whereby it contains computational units that mimic the structure and functions of neurons in the brain. The model has 8 encoder layers (left of Figure 1) and 8 decoder layers (right of Figure 1), with attention mechanisms between the layers.

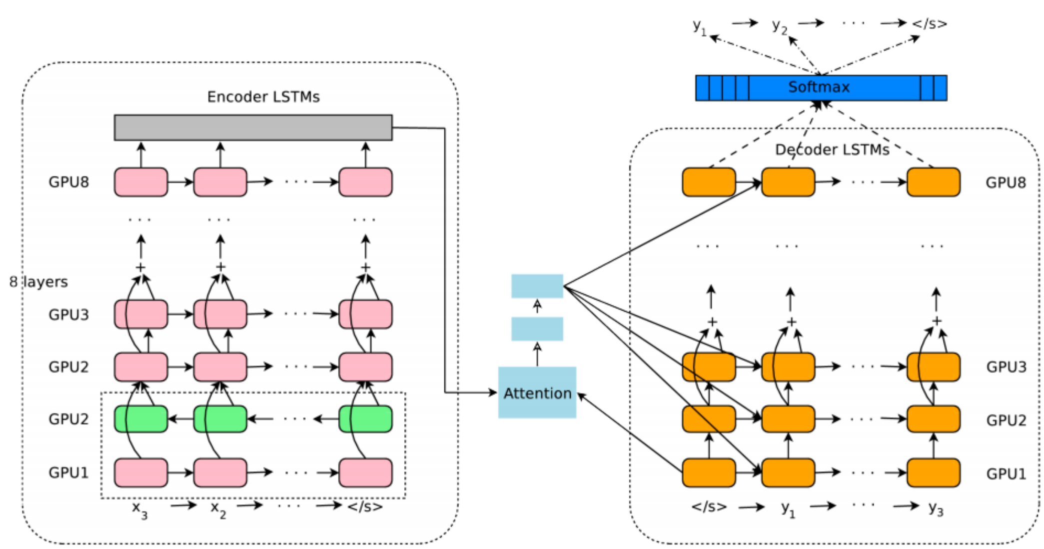

Figure 1: Model architecture of GNMT [2].

The encoder Recurrent Neural Network (RNN) is responsible for converting input symbols into vectors, which are representations of words in number form. Here, the conditional probabilities of word sequences are calculated. Recurrence allows the network to 'remember' and make use of all the preceding input information to calculate this probability. Google makes use of a bi-directional bottom encoder layer, which allows for words to be converted from both directions of sentences (beginning to end, and end to beginning) [3]. This means that the context of a single word within an entire sentence will be taken into account, thus enabling the algorithm to determine the conditional probabilities of word combinations in the sequence to a greater degree of accuracy.

An attention network joins the top encoder layer to the bottom decoder layer, suggesting words in the target language based on the vectors taken from the encoder [4]. It also allows for the decoder to 'pay attention' to different positions in the source sentence through its process of decoding [2]. The decoder RNN takes suggested words from the attention network and outputs them into sentences in the target language. As seen in Figure 1, a Softmax layer is combined with the decoder network and implemented just before the output layer. It assigns decimal probabilities to the candidate output symbols, which then determines the final output sentence [5], as shown in Figure 2.

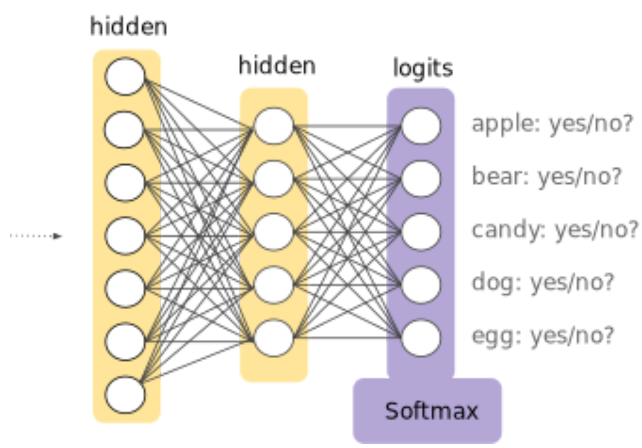

Figure 2: A Softmax layer determines the final output [5].

To train this system, Google utilises labelled datasets of parallel text derived from publicly available corpora: the 2014 Workshop on Machine Translation English-to-French (WMT En→Fr) and English-to-German (WMT En→De), as well as Google's own translation production database [2]. As a deep learning system, GNMT is able to identify differences and patterns between languages without the need for human input of such rules. This is especially useful, given the rules of languages are constantly changing. The reason for its multiple layers is that the more layers there are, the more the system is able to capture subtle differences between languages. These layers form a deep Long Short Term Memory (LSTM) network, a special kind of RNN that allows the system to choose to add or remove information to the output, while taking into consideration the preceding and succeeding information much more effortlessly than regular RNNs [6].

4 Limitations of Google Translate

Let us consider a case brought forward by Douglas Hofstadter, a professor of cognitive science and comparative literature, in an article from The Atlantic [7]: as he input a piece of Chinese text into the software, he realised that Google translate had incorrectly translated the phrase “南书房行走”. By character, the phrase literally translates to “south book room go walk”, and was translated by Google as “South study walking”. While Hofstadter did not know what this phrase meant, he could identify that it was supposed to be a noun instead of an action verb. As such, he used the Google search engine, through which he discovered that the two characters “行走" were rare terms used in the Qing Dynasty to mean “academic aide”. Hofstadter thus concluded that the original phrase could be translated as “South Study special aide”.

As shown, just because GNMT can identify patterns between languages and produce highly probable translations of certain words, does not mean that it can always follow the specific syntax of the target language. This affects the grammatical accuracy of output, and, as in the example above, their meaning.

5 Proposed Additions

Instead of relying solely on parallel texts, monolingual data should be incorporated into the training data for GNMT. By doing so, the system would be able to learn a language model for the target language, strengthening the syntax and meaning-accuracy of the output.

Monolingual data should be input into both the encoder and the decoder networks. Some research has suggested that systems tend to 'forget' information obtained from the encoder if trained on more monolingual data than parallel data [8]. As such, some proposals have been to have an encoder-independent layer that feeds information from the monolingual data into the decoder network, as done by Domhan & Hieber [8]. On the other hand, Senrich et al. have been successful in incorporating monolingual data into the encoder network, without having to change the architecture of their original neural network [9].

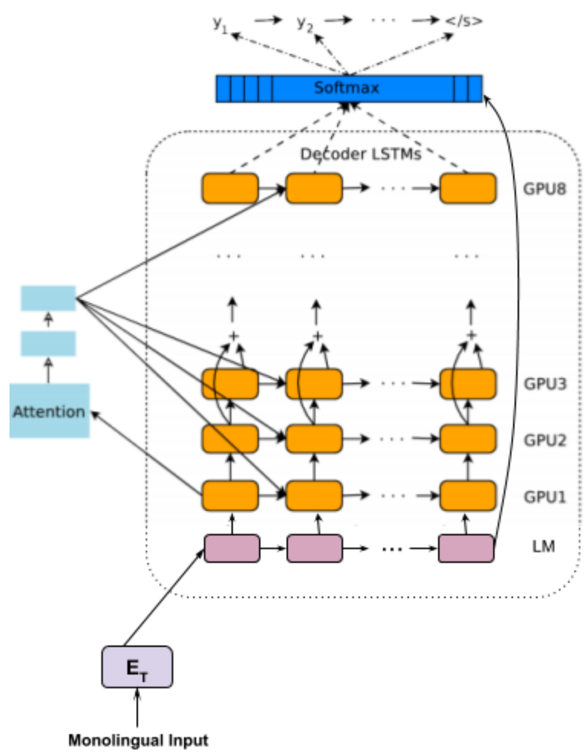

Figure 3: Proposed GNMT model (adapted from [2]). The encoder network is not shown and remains as in the original GNMT model [2].

Adapting from Domhan & Hieber's model [8], a new GNMT model is proposed as shown in Figure 3. An additional RNN Learning Model (LM) layer is connected to the decoder network. This LM layer is source-independent (is not connected to encoder), and only learns from a language model provided by ET (target word embedding matrix), which represents input words in the form of vectors, similar to the function of the encoder, as well as maps similar vectors closer together so as to allow the relations between words to be easily identified [10]. The LM layer thus uses the input from ET and predicts the next words in the target language, feeding this information to both the decoder RNN, to be taken into account together with predictions from the encoder RNN, as well as the Softmax layer, which produces an output based on its probability calculations. In order to ensure that systems do not ‘forget’ information from the encoder network, the proportion of monolingual data should be controlled to be less than 50% of the total training data.

The monolingual data should be derived from web pages, so as to keep the language model as up to date as possible, allowing for slangs and new vocabularies to be translatable. This may call for the need to train the system more frequently, for example, once a month, which is feasible given the training for GNMT takes roughly a week [2]. No architectural changes are made to the encoder network, however, changes need to be made to the input data. Senrich et al. employed the back-translation technique whereby the monolingual target-language text is translated back into the source language first, so as to create parallel text which the system can learn from [9].

Through these changes, instead of only translating inputs into probable outputs, the system may also learn to identify parts of speech before the translation process. Going back to Hofstadter’s case [7], it can be hoped that the GNMT system will learn to identify that the verb “做 (to do)" that appeared before “南书房行走” rendered the phrase a noun, and thus become able to suggest translations on the basis that it is a noun.

Furthermore, considering that much more monolingual data is available than parallel data [9], this model will prevent large amounts of monolingual data from going to waste.

6 Limitations

This proposed model remains a theoretical model given the scope of this paper, and has yet to be experimented on. At this moment, possible limitations that can be inferred are increases in training time, as well as contradictory decreases in the accuracy of output. The compilation of monolingual databases from web pages will also be time consuming and possibly costly, as the data needs to be quality-checked and labelled frequently.

7 Conclusion

Nevertheless, previous research suggests the viability of such a model, as well as its ability to result in more accurate, natural output. With this model, coupled with Google's recent update that allows users to review and contribute translations [11], Google Translate can be expected to become increasingly accurate in its translations. Further investigations could be how Google's Image Search results could be incorporated as training data, so as to help the system identify meanings of rare words. With Google Translate becoming ever closer to human translators, it is hoped that the software will not be seen as replacing human translators, but as augmenting their jobs, as well as increasing the level of everyday communication between people around the world.

References

[1] The Economist. *Finding a voice.* URL: https://www.economist.com/technology-quarterly/2017-05-01/language. (accessed: 26.11.2020).

[2] Yonghui Wu et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: (2016). arXiv: 1609.08144.

[3] Mike Schuster and Kuldip K. Paliwal. "Bidirectional Recurrent Neural Networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093.

[4] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *International Conference on Learning Representations* (2015). arXiv: 1409.0473.

[5] Google Developers. *Multi-Class Neural Networks: Softmax*. URL: https://developers.google .com/machine-learning/crash-course/multi-class-neural-networks/softmax. (accessed:10.01.2021).

[6] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[7] Douglas Hofstadter. *The Shallowness of Google Translate*. URL: https://www.theatlantic.com/ technology/archive/2018/01/the-shallowness-of-google-translate/551570/. (accessed: 26.11.2020).

[8] Tobias Domhan and Felix Hieber. "Using Target-side Monolingual Data for Neural Machine Translation through Multi-task Learning". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2017). DOI: 10.18653/v1/D17-1158:.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data". In: (2016). arXiv: 1511.06709v4.

[10] G. Salton, A. Wong, and C. S. Yang. "A Vector Space Model for Automatic Indexing". In: Communications of the ACM 18.11 (1975), pp. 613–620.

[11] Google Translate. *Improve Translate.* URL: https://translate.google.com/intl/en/about/ contribute/. (accessed: 11.1.2021).

Chapter 3

Submission 3: Lalejini

The benefits of self-replicating computer programs as model organisms for experimental evolution

Experimental evolution allows us to test general hypotheses about evolutionary processes by studying real-time evolutionary changes occurring in experimental populations (Kawecki et al., 2012). Indeed, experimental evolution is an essential technique in our methodological repertoire for understanding the evolution of life as it is, life as it could have been, and life as it might be. Conventionally, evolution experiments are performed under laboratory conditions using populations of biological organisms that are tractable to observe and experimentally manipulate (e.g., *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and various phage-bacteria systems). For example, over 70,000 generations of evolution have elapsed in the ongoing long-term evolution experiment with *E. coli* (Barrick et al., 2020), which has yielded insights on a wide range of topics, including long-term evolutionary dynamics (Wiser et al., 2013; Good et al., 2017), historical contingency (Travisano et al., 1995; Card et al., 2019), the evolution of mutation rates (Sniegowski et al., 1997), and the origins of novel traits (Blount et al., 2008). Here, I discuss the benefits of using artificial life systems to conduct evolution experiments; in particular, I review the value of self-replicating computer programs as model organisms for experimental evolution.

Digital evolution experiments have emerged as a powerful research framework from which evolution can be studied. In digital evolution, self-replicating computer programs (digital organisms) compete for resources, mutate, and evolve in a computational environment (Wilke and Adami, 2002). Digital organisms typically comprise a linear sequence of program instructions (a genome) and a set of virtual hardware components used to interpret and express those instructions. To reproduce, a digital organism must execute instructions that allow it to copy its genome instruction-by-instruction and then divide (producing an offspring). However, self-replication is imperfect and can result in mutated offspring. The combination of heritable variation due to imperfect self-replication and competition for limited resources (e.g., space, CPU time, *etc.*) results in evolution by natural selection.

Digital organisms live, interact, and evolve in entirely artificial environments constructed by the experimenters. One potential drawback to digital evolution is that the conclusions drawn from an experiment have the potential to be artifacts of the constructed artificial environment (Wilke and Adami, 2002). This drawback, however, can also be applied to most microbial experimental evolution where organisms are extracted from their natural environment and placed in an artificial environment constructed in a laboratory.

Microbial model organisms at least have natural ancestry and can often be used to infer historic evolutionary events. Digital evolution studies, however, are not grounded in the same evolutionary history and biochemical compounds as carbon-based life on Earth. This limitation makes it more challenging to use digital evolution studies to illuminate idiosyncrasies and contingencies associated with the history of life on our planet. However, such drawbacks are also digital evolution's strength as a research framework, since we are not limited to studying only one particular instance of evolution or locked in to using nucleic-acid, amino acid, and protein based representations. Furthermore, we can fully observe and control digital environments at rapid speeds, allowing us to perform experiments and analyses that would otherwise be challenging or even impossible to perform in biological systems. Additionally, by reproducing results across biological and digital systems, we can disentangle general principles from effects specific to a particular model organism or planetary body (Wilke and Adami, 2002).

Here, I overview four properties of digital evolution systems that make them valuable complements to traditional carbon-based model organisms for studying evolutionary processes, providing exemplars of each: (1) generality, (2) transparency, (3) control, and (4) scale.

Generality

Digital evolution systems offer researchers the unique opportunity to study evolution in organisms that share no ancestry with carbon-based life (Wilke and Adami, 2002). As biologist John Maynard Smith made the case, "So far, we have been able to study only one evolving system and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we will have to look to artificial ones" (Maynard Smith, 1992). Indeed, studies of carbon-based lifeforms that all share common ancestry dominate evolutionary biology. On their own, these studies can provide deeper insights into life on Earth. However, such studies provide a limited lens with which to make generalizations about evolutionary processes, as they are biased by the particular history of life on our planet. By testing hypotheses across biological and digital model systems, we can disentangle general principles from the effects of specific model organisms.

For example, what is the relative importance of adaptation, chance, and history in explaining diversity in evolved populations? Using experimental populations of *Escherichia coli*, Travisano et al. disentangled the relative contributions of adaptation, chance, and history in the evolution of fitness and cell size (a trait weakly correlated with fitness) (Travisano et al., 1995).

Travisano et al. found that fitness gains were most strongly influenced by adaptive processes, and variance in cell size were most explained by chance and history. Wagenaar and Adami replicated this study using digital organisms (Wagenaar and Adami, 2004), finding that the overall patterns observed in *E. coli* and in digital organisms were broadly similar. Ongoing studies in digital organisms are extending these concepts further, using more restarts at different time points and across different environments, allowing us to explore more of the nuances at play (Bundy et al., 2021).

Transparency

Digital evolution systems allow for perfect, non-invasive data tracking. Experimenters can save the complete details of evolving populations for further analysis, including every mutation that occurs, every genotype that exists, every phenotype that is expressed, every environmental state that occurs, every time an organism interacts with another organism or with the environment, *et cetera*. By tracking parent-offspring relationships, we can analyze complete evolutionary histories within an experiment, which circumvents the historical problem of drawing evolutionary inferences using incomplete records (from frozen samples or fossils) and extant genetic sequences.

Many digital evolution studies inspect the complete lineages of evolved digital organisms to tease apart the mutation-by-mutation evolution of novel traits (Lenski et al., 2003; Dolson and Ofria, 2017; Grabowski et al., 2013; Goldsby et al., 2014; Pontes et al., 2020). In an exemplary analytical undertaking, Dolson and Ofria identified spatial hotspots of evolutionary potential in heterogeneous environments (i.e., positions where novel traits disproportionately evolved). They found evidence that the particular *paths* traversed by lineages through space might explain the locations of these evolutionary hotspots (Dolson and Ofria, 2017).

Recording organism relationships and interactions can be valuable for many other goals as well. For example, by tracking phenotypes over time, Cooper and Ofria were able to observe the real-time evolution of stable ecosystems under resource-limited conditions (Cooper and Ofria, 2002). In a similar vein, Fortuna et al. tracked host-parasite interactions to investigate how the structure of infection networks is shaped by antagonistic coevolution (Fortuna et al., 2019).

Control

Digital evolution systems facilitate experimental manipulations that go beyond what is possible in laboratory or field experiments. These capabilities allow researchers to empirically test hypotheses that would otherwise be relegated to theoretical analyses. For example, digital evolution systems allow experimenters to precisely control basic parameters such as population size and mutation rate. By comparing populations evolving under different mutation rates, Wilke et al. discovered the "survival of the flattest" effect where high mutation rate environments selected for genomes with slower replication rates but that were more robust to mutations (Wilke et al., 2001).

Digital evolution experiments also allow for fine-grained control over other aspects of an environment. For example, Dolson et al. used Avida to experimentally manipulate the spatial distribution of resource availability, finding that phenotypic diversity was positively correlated with spatial entropy and that spatially heterogeneous environments exhibited increased evolutionary potential relative to more homogeneous environments (Dolson et al., 2017). By experimentally controlling how environments changed temporally, Nahum et al. demonstrated that a single temporary environmental change can improve fitness landscape exploration and exploitation in evolving populations of digital organisms (Nahum et al., 2017).

Digital evolution systems also allow experimenters to monitor and manipulate mutational effects in real-time. Covert et al. performed real-time reversions of all deleterious mutations as they occurred to isolate their long-term effects on evolutionary outcomes (Covert et al., 2013). Lalejini et al. implemented a range of gene duplication mutation operators (each designed to isolate a single effect of duplication mutations) in order to tease apart why such mutations can promote the evolution of complex traits (Lalejini et al., 2017).

For an individual digital organism, we can perform systematic knockout analyses to identify which instructions are responsible for producing a given phenotypic outcome. This type of analysis has been applied along lineages to identify how information accumulates (Ofria et al., 2008) or to investigate how environmental change shapes the evolution of genetic architectures in digital organisms (Canino-Koning et al., 2016). Mutational landscaping analyses go a step further than knockout analyses, allowing experimenters to fully characterize a local mutational landscape by evaluating all possible one- and two-step mutants.
Such analyses have been used to quantify epistasis (Lenski et al., 1999) and mutational robustness (Elena et al., 2007) and to investigate the evolution of evolvability (Canino-Koning et al., 2019).

Scale

Modern computers allow us to observe many generations of digital evolution at tractable time scales; thousands of generations can take mere minutes as opposed to months, years, or centuries. For example, populations of digital organisms have been used to test theoretical predictions about the expected rate of adaptation over hundreds of thousands of generations (Wiser 2015; Wiser et al. 2018).

Additionally, digital evolution experiments allow researchers to enact complex experimental protocols with minimal extra effort. That is, unlike in wet-lab experiments, computational experiment protocols can easily be automated using modern scripting tools.

With the increasing accessibility of high performance computing systems, it can be trivial to evolve hundreds of replicate populations for a given experimental treatment. Evolution is an inherently stochastic process, so increased replication provides a clearer picture of the distribution of possible treatment effects. Further, a high degree of replication increases the odds that experimenters will be able to observe and study rare events. For example, Pontes et al. evolved 900 replicate populations of digital organisms in order to observe 10 examples of reversal learning behavior (i.e., the ability to relearn associations between cues and responses when cues are swapped) to further analyze (Pontes et al., 2020).

Even the fastest computing systems, however, lack the parallelism of the real world.
That is, digital evolution systems cannot yet rival bacterial systems in their ability to scale to large population sizes. A typical population of digital organisms contains thousands to tens of

thousands of organisms; however, microbial populations used in laboratory experiments often contain several orders of magnitude more individuals.

Outlook

In this brief review, I highlighted four features of evolution experiments with digital organisms—generality, transparency, control, and scale—that make digital experiments a valuable methodological complement to more conventional evolution experiments with carbon-based model organisms. While digital organisms have a proven track record as a model system, ongoing research continues to advance the applicability of digital organisms for studying the general principles of evolution. For example, there are many different model organisms used in biological research, each with their own benefits and shortcomings for conducting evolution experiments; yet, historically, there have been very few different forms of self-replicating computer programs used in digital evolution experiments. As such, new ways of representing and interpreting digital organisms are being developed (e.g., Lalejini and Ofria, 2018). Additionally, digital evolution researchers are actively exploring new types of digital environments (e.g., Moreno and Ofria, 2021), analyses (e.g., Dolson et al., 2020), and visualizations (e.g., Dolson and Ofria, 2018), each of which expand the set of investigations possible with digital organisms. Finally, new hardware (Ackley, 2020) and software (Ackley and Cannon, 2011; Moreno and Ofria, 2020) systems designed with artificial life in mind will continue to increase the scale of digital evolution experiments beyond our current limitations.

References

References abbreviated here to save space. Full references provided in a supplemental file. If this work is posted somewhere, please contact me for a version that includes the full reference list.

Ackley, D. H. (2020). https://doi.org/10.1162/isal_a_00337

Ackley, D. H., & Cannon, D. C. (2011).https://dl.acm.org/doi/10.5555/1991596.1991607

Barrick, J. E., et al. (2020). https://doi.org/10.1007/978-3-030-39831-6_8

Blount, Z. D., et al. (2008). https://doi.org/10.1073/pnas.0803151105

Bundy, J. N., et al. (2021). https://doi.org/10.1101/2021.04.29.442046

Canino-Koning, R., et al. (2019). https://doi.org/10.1371/journal.pcbi.1006445

Canino-Koning, R. (2016). https://doi.org/10.7551/978-0-262-33936-0-ch047

Card, K. J., et al. (2019). https://doi.org/10.1371/journal.pbio.3000397

Cooper, T. F., & Ofria, C. (2002). https://dl.acm.org/doi/10.5555/860295.860334

Covert, A. W., et al. (2013). https://doi.org/10.1073/pnas.1313424110

Dolson, E., et al. (2017). https://doi.org/10.1101/148973

Dolson, E., et al. (2020). https://doi.org/10.1162/artl_a_00313

Dolson, E., & Ofria, C. (2017). https://doi.org/10.7551/ecal_a_023

Dolson, E., & Ofria, C. (2018). https://doi.org/10.1145/3205651.3208301

Elena, S. F., et al. (2007). https://doi.org/10.1111/j.1558-5646.2007.00064.x

Fortuna, M. A., et al. (2019). https://doi.org/10.1111/evo.13731

Goldsby, H. J., et al. (2014). https://doi.org/10.1371/journal.pbio.1001858

Good, B. H., et al.(2017). https://doi.org/10.1038/nature24287

Grabowski, L. M., et al. (2013). https://doi.org/10.1371/journal.pone.0060466

Kawecki, T. J., et al. (2012). https://doi.org/10.1016/j.tree.2012.06.001

Lalejini, A., & Ofria, C. (2018). https://doi.org/10.1162/isal_a_00069

Lalejini, A., et al. (2017). https://doi.org/10.7551/ecal_a_045

Lenski, R. E., et al. (1999). https://doi.org/10.1038/23245

Lenski, R. E., et al. (2003). https://doi.org/10.1038/nature01568

Maynard Smith, J. (1992). https://doi.org/10.1038/355772a0

Moreno, M. A, & Ofria, C. (2020). https://doi.org/10.17605/OSF.IO/53VGH

Moreno, M. A., & Ofria, C. (2021). https://arxiv.org/abs/2104.10081

Nahum, J. R., et al. (2017). https://doi.org/10.7551/ecal_a_052

Ofria, C., et al. (2008). https://doi.org/10.1162/artl.2008.14.3.14302

Pontes, A. C., et al. (2020). https://doi.org/10.1086/706252

Sniegowski, P. D., et al. (1997). https://doi.org/10.1038/42701

Travisano, M., et al. (1995). https://doi.org/10.1126/science.7809610

Wagenaar, D. A., & Adami, C. (2004). https://doi.org/10.1162/106454604773563603

Wilke, C. O., & Adami, C. (2002). https://doi.org/10.1016/S0169-5347(02)02612-5

Wilke, C. O., et al. (2001). https://doi.org/10.1038/35085569

Wiser, M. J., et al. (2013). https://doi.org/10.1126/science.1243357

Wiser, M. J. (2015). PhD thesis.

Wiser, M. J., et al. (2018). https://doi.org/10.7287/peerj.preprints.27246v2

Chapter 4

Submission 4: Moreno

Scalability without Isolation is Critical for Future Digital Evolution Models

Digital evolution techniques compliment traditional wet-lab evolution experiments by enabling researchers to address questions that would be otherwise limited by:

- reproduction rate (which determines the number of generations that can be observed in a set amount of time),
- incomplete observations (every event in a digital system can be tracked),
- physically-impossible experimental manipulations (every event in a digital system can can be arbitrarily altered), or
- resource- and labor-intensity (digital experiments and assays can be easily automated).

The versatility and rapid generational turnover of digital systems can easily engender a notion that such systems can already operate at scales greatly exceeding biological evolution experiments. Although digital evolution techniques can feasibly simulate populations numbering in the millions or billions, very simple agents and/or very limited agent-agent interaction. With more complex agents controlled by genetic programs, neural networks, or the like, feasible population sizes dwindle down to thousands or hundreds of agents.

Putting Scale in Perspective

Take Avida as an example. This popular software system that enables experiments with evolving self-replicating computer programs. In this system, a population of ten thousand can undergo about twenty thousand generations per day. This means that about two hundred million replication cycles are performed in a day [Ofria et al., 2009].

Each flask in the Lenski Long-Term Evolution Experiment hosts a similar number of replication cycles. In their system, E. coli undergo about six doublings per day. Effective population size is reported as 30 million [Good et al., 2017]. Hence, about 180 million replication cycles elapse per day.

Likewise, in Ratcliff's work studying the evolution of multicellularity in S. cerevisiae, about six doublings per day occur among a population numbering on the order of a billion cells [Ratcliff, 2012]. So, around six billion cellular replication cycles elapse per day in this system.

Although artificial life practitioners traditionally describe instances of their simulations as "worlds," with serial processing power their scale aligns (in naive terms) more along the lines of a single flask. Of course, such a comparison neglects the disparity between Avidians and bacteria or yeast in terms of genome information content, information content of cellular state, and both quantity and diversity of interactions with the environment and with other cells.

Recent work with SignalGP has sought to address some of these shortcomings by developing digital evolution substrates suited to more dynamic environmental and agent-agent interactions [Lalejini and Ofria, 2018] that more effectively incorporate state information [Lalejini et al., 2020; Moreno, 2020]. However, to some degree, more sophisticated and interactive evolving agents will necessarily consume more CPU time on a per-replication-cycle basis — further shrinking the magnitude of experiments tractable with serial processing.

The Future is Parallel

Throughout the 20th century, serial processing enjoyed regular advances in computational capacity due to quickening clock cycles, burgeoning RAM caches, and increasingly clever packing together of instructions during execution. Since, however, performance of serial processing has bumped up against apparent fundamental limits to computing's current technological incarnation [Sutter, 2005]. Instead, advances in 21st century computing power have arrived via multiprocessing [Hennessy and Patterson, 2011, p.55] and hardware acceleration (e.g., GPU, FPGA, etc.) [Che et al., 2008].

Contemporary high-performance computing clusters link multiprocessors and accelerators with fast interconnects to enable coordinated work on a single problem [Hennessy and Patterson, 2011, p.436]. High-end clusters already make hundreds of thousands or millions of cores available. More loosely-affiliated banks of servers can also muster significant computational power. For example, Sentient Technologies notably employed a distributed network of over a million CPUs to run evolutionary algorithms [Miikkulainen et al., 2019].

The availability of orders of magnitude greater parallel computing resources in ten and twenty years' time seems probable, whether through incremental advances with traditional silicon-based technology or via emerging, unconventional technologies such as bio-computing [Benenson, 2009] and molecular electronics [[Xiang et al., 2016]](#xiang2016molecular. Such emerging technologies could make greatly vaster collections of computing devices feasible, albeit at the potential cost of component-wise speed [Bonnet et al., 2013](#bonnet2013amplifying); Ellenbogen and Love, 2000] and perhaps also component-wise reliability.

<u>What of Scale?</u>

Digital evolution practitioners have a rich history of leveraging distributed hardware. It is common practice to distribute multiple self-isolated instantiations of evolutionary runs over multiple hardware units. In scientific contexts, this practice yields replicate datasets that provide statistical power to answer research questions [Dolson and Ofria, 2017]. In applied contexts, this practice yields many converged populations that can be scavenged for the best solutions overall [Hornby et al., 2006].

Another established practice is to use “island models” where individuals are transplanted between populations that are otherwise independently evolving across distributed hardware. Koza and collaborators’ genetic programming work with a 1,000-cpu Beowulf cluster typifies this approach [Bennett III et al., 1999].

In recent years, Sentient Technologies spearheaded digital evolution projects on an unprecedented computational scale, comprising over a million CPUs and capable of a peak performance of 9 petaflops [Miikkulainen et al., 2019]. According to its proponents, the scale and scalability of this DarkCycle system was a key aspect of its conceptualization [Gilbert, 2015]. Much of the assembled infrastructure was pieced together from heterogeneous providers and employed on a time-available basis [Blondeau et al., 2012]. Unlike island model where selection events are performed independently on each CPU, this scheme transferred evaluation criteria between computational instances (in addition to individual genomes) [Hodjat and Shahrzad, 2013].

Sentient Technologies also accelerated the deep learning training process by using many massively-parallel hardware accelerators (e.g., 100 GPUs) to evaluate the performance of candidate neural network architectures on image classification, language modeling, and image captioning problems [Miikkulainen et al., 2019]. Analogous work parallelizing the evaluation of an evolutionary individual over multiple test cases in the context of genetic programming has used GPU hardware and vectorized CPU operations [Harding and Banzhaf, 2007b; Langdon and Banzhaf, 2019].

Existing applications of concurrent approaches to digital evolution distribute populations or individuals across hardware to process them with minimal interaction. Task independence facilitates this simple, efficient implementation strategy, but precludes application on elements that are not independent. Parallelizing evaluation of a single individual often emphasizes data-parallelism over independent test cases, which are subsequently consolidated into a single fitness profile. With respect to model parallelism, Harding has notably applied GPU acceleration to cellular automata models of artificial development systems, which involve intensive interaction between spatially-distributed instantiation of a genetic program [Harding and Banzhaf, 2007a]. However, in systems where evolutionary individuals themselves are parallelized they are typically completely isolated from each other.

We argue that, in a manner explicitly accommodating capabilities and limitations of available hardware, open-ended evolution should prioritize dynamic interactions between simulation elements situated across physically distributed hardware components.

Leveraging Distributed Hardware for Open-Ended Evolution

Unlike most existing applications of distributed computing in digital evolution, open-ended evolution researchers should prioritize dynamic interactions among distributed simulation elements. Parallel and distributed computing enables larger populations and metapopulations. However, ecologies, co-evolutionary dynamics, and social behavior all necessitate dynamic interactions among individuals.

Distributed computing should also enable more computationally intensive or complex individuals. Developmental processes and emergent functionality necessitate dynamic interactions among components of an evolving individual. Even at a scale where individuals remain computationally tractable on a single hardware component, modeling them as a collection of discrete components configured through generative development (i.e., with indirect genetic representation) can promote scalable properties [Lipson, 2007] such as modularity, regularity, and hierarchy [Hornby, 2005; Clune et al., 2011]. Developmental processes may also promote canalization [Stanley and Miikkulainen, 2003], for example through exploratory processes and compensatory adjustments [Gerhart and Kirschner, 2007]. To reach this goal, David Ackley has envisioned an ambitious design for modular distributed hardware at a theoretically unlimited scale [Ackley and Cannon, 2011] and demonstrated an algorithmic substrate for emergent agents that can take advantage of it [Ackley, 2018].

A Path of Expanding Computational Scale

While by no means certain, the idea that orders-of-magnitude increases in compute power will open up qualitatively different possibilities with respect to open-ended evolution is well founded. Spectacular advances achieved with artificial neural networks over the last decade illuminate a possible path toward this outcome. As with digital evolution, artificial neural networks (ANNs) were traditionally understood as a versatile, but auxiliary methodology — both techniques were described as "the second best way to do almost anything" [Miaoulis and Plemenos, 2008; Eiben, 2015]. However, the utility and ubiquity of ANNs has since increased dramatically. The development of AlexNet is widely considered pivotal to this transformation. AlexNet united methodological innovations from the field (such as big datasets, dropout, and ReLU) with GPU computing that enabled training of orders-of-magnitude-larger networks. In fact, some aspects of their deep learning architecture were expressly modified to accommodate multi-GPU training [Krizhevsky et al., 2012]. By adapting existing methodology to exploit commercially available hardware, AlexNet spurred the greater availability of compute resources to the research domain and eventually the introduction of custom hardware to expressly support deep learning [Jouppi et al., 2017].

Similarly, progress toward realizing artificial life systems with indefinite scalability seems likely to unfold as incremental achievements that spur additional interest and resources in a positive feedback loop with the development of methodology, software, and eventually specialized hardware to take advantage of those resources. In addition to developing hardware-agnostic theory and methodology, we believe that pushing the envelope of open-ended evolution will analogously require designing systems that leverage existing commercially-available parallel and distributed compute resources at circumstantially-feasible scales.

References

Ackley, D. H. (2018). Digital protocells with dynamic size, position, and topology. The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE), pages 83–90.

Ackley, D. H. and Cannon, D. C. (2011). Pursue robust indefinite scalability. In HotOS.

Bassett, D. S. and Bullmore, E. T. (2017). Small-world brain networks revisited. The Neuroscientist, 23(5):499–516.

Bedau, M. A., Snyder, E., and Packard, N. H. (1998). A classification of long-term evolutionary dynamics. In Artificial life VI, pages 228–237.

Benenson, Y. (2009). Biocomputers: from test tubes to live cells. Molecular BioSystems, 5(7), 675-685.

Bennett III, F. H., Koza, J. R., Shipman, J., and Stiffelman, O. (1999). Building a parallel computer system for $18,000 that performs a half peta-flop per day. In Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 2, pages 1484–1490. Citeseer.

Blondeau, A., Cheyer, A., Hodjat, B., and Harrigan, P. (2012). Distributed network for performing complex algorithms. US Patent App. 13/443,546.

Bohm, C. and Hintze, A. (2017). Mabe (modular agent based evolver): A framework for digital evolution research. In Artificial Life Conference Proceedings 14, pages 76–83. MIT Press.

Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P., & Endy, D. (2013). Amplifying genetic logic gates. Science, 340(6132), 599-603.

Channon, A. (2019). Maximum individual complexity is indefinitely scalable in geb. Artificial life, 25(2):134–144.

Che, S., Li, J., Sheaffer, J. W., Skadron, K., and Lach, J. (2008). Accelerating compute-intensive applications with gpus and fpgas. In 2008 Symposium on Application Specific Processors, pages 101–107. IEEE.

Clarke, L., Glendinning, I., and Hempel, R. (1994). The mpi message passing interface standard. In Programming environments for massively parallel distributed systems, pages 213–218. Springer.

Clune, J., Stanley, K. O., Pennock, R. T., and Ofria, C. (2011). On the performance of indirect encoding across the continuum of regularity. IEEE Transactions on Evolutionary Computation, 15(3):346–367.

Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. IEEE computational science and engineering, 5(1):46–55.

Dolson, E. and Ofria, C. (2017). Spatial resource heterogeneity creates local hotspots of evolutionary potential. In Artificial Life Conference Proceedings 14, pages 122–129. MIT Press.

Dolson, E. L. (2019). On the Constructive Power of Ecology in Open-Ended Evolving Systems. Michigan State University.

Dolson, E. L., Vostinar, A. E., Wiser, M. J., and Ofria, C. (2019). The modes toolbox: Measurements of open-ended dynamics in evolving systems. Artificial life, 25(1):50–73.

Eiben, A. and Smith, J. E. (2015). Introduction to evolutionary computing. Springer, Berlin.

Ellenbogen, J. C., & Love, J. C. (2000). Architectures for molecular electronic computers. I. Logic structures and an adder designed from molecular electronic diodes. Proceedings of the IEEE, 88(3), 386-426.

Foster, E. D. and Deardorff, A. (2017). Open science framework (osf). Journal of the Medical Library Association: JMLA, 105(2):203.

Fox, R. J. and Bellwood, D. R. (2014). Herbivores in a small world: network theory highlights vulnerability in the function of herbivory on coral reefs. Functional Ecology, 28(3):642–651.

Gaiteri, C., Ding, Y., French, B., Tseng, G. C., and Sibille, E. (2014). Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes, brain and behavior, 13(1):13–24.

Gerhart, J. and Kirschner, M. (2007). The theory of facilitated variation. Proceedings of the National Academy of Sciences, 104(suppl 1):8582–8589.

Gilbert, D. (2015). Artificial intelligence is here to help you pick the right shoes.

Goldsby, H. J., Dornhaus, A., Kerr, B., and Ofria, C. (2012). Taskswitching costs promote the evolution of division of labor and shifts in individuality. Proceedings of the National Academy of Sciences, 109(34):13686–13691.

Goldsby, H. J., Knoester, D. B., Ofria, C., and Kerr, B. (2014). The evolutionary origin of somatic cells under the dirty work hypothesis. PLOS Biology, 12(5):e1001858.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., & Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. Nature, 551(7678), 45–50.

Harding, S. and Banzhaf, W. (2007a). Fast genetic programming and artificial developmental systems on gpus. In 21st International Symposium on High Performance Computing Systems and Applications (HPCS'07), pages 2–2. IEEE.

Harding, S. and Banzhaf, W. (2007b). Fast genetic programming on gpus. In European conference on genetic programming, pages 90–101. Springer.

Hennessy, J. L. and Patterson, D. A. (2011). Computer architecture: a quantitative approach. Elsevier.

Hodjat, B. and Shahrzad, H. (2013). Distributed evolutionary algorithm for asset management and trading. US Patent 8,527,433.

Hornby, G. S. (2005). Measuring, enabling and comparing modularity, regularity and hierarchy in evolutionary design. In Proceedings of the 7th annual conference on Genetic and evolutionary computation, pages 1729–1736.

Hornby, G., Globus, A., Linden, D., and Lohn, J. (2006). Automated antenna design with evolutionary algorithms. Space 2006.

Huizinga, J., Stanley, K. O., and Clune, J. (2018). The emergence of canalization and evolvability in an open-ended, interactive evolutionary system. Artificial life, 24(3):157–181.

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., et al. (2017). In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th Annual International Symposium on Computer Architecture, pages 1–12.

Kirschner, M. and Gerhart, J. (1998). Evolvability. Proceedings of the National Academy of Sciences, 95(15):8420–8427.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105.

Kuffner, J. J. and LaValle, S. M. (2009). Space-filling trees. RI, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-09-47

Lalejini, A. and Ofria, C. (2018). Evolving event-driven programs with signalgp. In Proceedings of the Genetic and Evolutionary Computation Conference, pages 1135–1142.

Lalejini, A., Moreno, M. A., & Ofria, C. (2020, June 27). Case Study of Adaptive Gene Regulation in DISHTINY.

Langdon, W. B. and Banzhaf, W. (2019). Continuous long-term evolution of genetic programming. In The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE), pages 388–395. MIT Press.

Legenstein, R. A. and Maass, W. (2001). Optimizing the layout of a balanced tree. In Electronic Colloquium on Computational Complexity (ECCC), volume 8.

Lehman, J. (2012). Evolution through the search for novelty.

Lehman, J. and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. Evolutionary computation, 19(2):189–223.

Lehman, J. and Stanley, K. O. (2012). Beyond open-endedness: Quantifying impressiveness. In Artificial Life Conference Proceedings 12, pages 75–82. MIT Press.

Lehman, J. and Stanley, K. O. (2013). Evolvability is inevitable: Increasing evolvability without the pressure to adapt. PloS one, 8(4).

Lipson, H. (2007). Principles of modularity, regularity, and hierarchy for scalable systems. Journal of Biological Physics and Chemistry, 7(4):125–128.

Lu, L. and Zhang, M. (2013). Edge Betweenness Centrality, pages 647–648. Springer New York, New York, NY.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences, 104(suppl 1):8597–8604.

Markov, I. L. (2014). Limits on fundamental limits to computation. Nature, 512(7513):147–154.

Miaoulis, G. and Plemenos, D. (2008). Intelligent Scene Modelling Information Systems, volume 181. Springer.

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., et al. (2019). Evolving deep neural networks. In Artificial Intelligence in the Age of Neural Networks and Brain Computing, pages 293–312. Elsevier.

Mollah, M. A., Faizian, P., Rahman, M. S., Yuan, X., Pakin, S., and Lang, M. (2018). A comparative study of topology design approaches for hpc interconnects. In 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pages 392–401. IEEE.

Moreno, M. A. (2020, July 8). Estimating DISHTINY Multithread Scaling Properties.

Moreno, M. A. (2020, May 1). Evaluating Function Dispatch Methods in SignalGP.

Moreno, M. A. and Ofria, C. (2019). Toward open-ended fraternal transitions in individuality. Artificial life, 25(2):117–133.

Moreno, M. A. and Ofria, C. (2020). Practical steps toward indefinite scalability: In pursuit of robust computational substrates for open-ended evolution. DOI: 10.17605/OSF.IO/53VGH; URL: https://osf.io/53vgh.

Moreno, M. A. and Ofria, C. (in prep.). Spatial constraints and kin recognition can produce open-ended major evolutionary transitions in a digital evolution system. https://doi.org/10.17605/OSF.IO/G58XK.

Moreno, M. A., & Ofria, C. (2020, June 25). Practical Steps Toward Indefinite Scalability: In Pursuit of Robust Computational Substrates for Open-Ended Evolution.

Nguyen, A. M., Yosinski, J., and Clune, J. (2015). Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, pages 959–966.

Ofria C., Bryson D.M., Wilke C.O. (2009) Avida. In: Komosinski M., Adamatzky A. (eds) Artificial Life Models in Software. Springer, London

Ofria, C., Dolson, E., Lalejini, A., Fenton, J., Moreno, M. A., Jorgensen, S., Miller, R., Stredwick, J., Zaman, L., Schossau, J., Gillespie, L., G, N. C., and Vostinar, A. (2019). Empirical.

Packard, N., Bedau, M. A., Channon, A., Ikegami, T., Rasmussen, S., Stanley, K. O., and Taylor, T. (2019). An overview of open-ended evolution: Editorial introduction to the open-ended evolution ii special issue. Artificial life, 25(2):93–103.

Ratcliff, W., Denison, R., Borrello, M., & Travisano, M. (2012). Experimental evolution of multicellularity. Proceedings of the National Academy of Sciences, 109(5), 1595–1600.

Soros, L. and Stanley, K. (2014). Identifying necessary conditions for open-ended evolution through the artificial life world of chromaria. In Artificial Life Conference Proceedings 14, pages 793–800. MIT Press.

Stanley, K. O. and Miikkulainen, R. (2003). A taxonomy for artificial embryogeny. Artificial Life, 9(2):93–130.

Stanley, K. O., Lehman, J., and Soros, L. (2017). Open-endedness: The last grand challenge you’ve never heard of. O’Reilly Online.

Sutter, H. (2005). The free lunch is over: A fundamental turn toward concurrency in software. Dr. Dobb’s journal, 30(3):202–210.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393(6684):440.

Xiang, D., Wang, X., Jia, C., Lee, T., & Guo, X. (2016). Molecular-scale electronics: from concept to function. Chemical reviews, 116(7), 4318-4440.

Chapter 5

Submission 5: Skocelas

The Evolution of Restraint on Multicellularity in Avida 4

1 Introduction

Artificial selection has been used by evolution researchers to study and explain the principle of natural selection for over one hundred years. For example, in *The Origin of Species*, Darwin (1859) used examples of artificial selection to lay the groundwork for his natural selection analog [1]. Though debate continues about the role of group selection in natural settings [2], more recent experiments using artificial selection have significantly increased our understanding of short- and long-term evolutionary processes [3].

One type of group selection, called antagonistic multilevel selection, is frequently seen in natural groups that exhibit division of labor [1]. In this type of selection, the between-group pressure to increase the group's fitness is at partial or total odds with the within-group pressure for individuals within the group to increase their fitness. For example, the cells of a multicellular organism undergo a within-group pressure to specialize on the role with the highest reward and a between-group pressure for the organism's cells to perform a diverse suite of tasks [4]. When the within-group pressure for cells to reproduce (i.e., go rogue to increase their own fitness) overwhelms the between-group pressure to control cellular reproduction, it results in cancer [5].

In this paper, we use the Avida digital evolution platform [6] to address the question: Do multicells evolve restraints on multicellularity when under treatments that allow rogue cell behavior? Digital evolution allows us to evolve two groups of digital multicells – a control group in which rogue cell behavior is programmatically prohibited, and a treatment group in which it is allowed. First, we look for evidence of cancer-like behavior in the treatment group. Second, we look for a decline over time in the amount of cancer-like behavior exhibited by a multicell. If restraints on multicellularity do not evolve in the treatment group, then it will not show a decline in the amount of cancer-like behavior exhibited. If cancer-like behavior emerges in the treatment group and then diminishes, however, natural selection for cancer prevention has evolved.

2 Background

2.1 Major Transitions Research with Avida 4

This project builds on Devolab alumni Dr. Heather Goldsby's work using Avida 4 to study the major evolutionary transition from single to multicell organisms. Avida is an open-source scientific software platform for conducting and analyzing experiments with self-replicating and evolving computer programs [6]. In 2012, Goldsby, Dornhaus, Kerr, and Ofria found that task-switching costs promote the evolution of division of labor and shifts in individuality between cells in digital multicellular organisms in Avida 4 [7]. Their study showed that higher task switching costs lead to increased division of labor among cells.

Next, Goldsby, Kerr, Ofria, and Knoester, proposed the "dirty work hypothesis" as the evolutionary origin of somatic cells, which stated that mutagenic effects associated with metabolism promote the evolution of germ-soma differentiation in multicells [8]. The study found that psuedo-soma cells – precursors to reproductive division of labor – do more mutagenic work, while the multicells preserved their genetic material by having a subset of cells do less or no mutagenic work. In this way, multicells made use of phenotypic plasticity to divide the workload prior to the evolution of somatic cells [8].

In the same year, Goldsby, Knoester, Kerr and Ofria looked at the effects of conflicting pressures on the evolution of division of labor, exploring how populations respond to antagonistic multilevel selection pressures [1]. They found that digital organisms from lineages performing highly rewarded roles used reproductive restraint in order to co-existence with organisms from other lineages. This inspired the question: Could multicells evolve similar forms of reproductive restraint to suppress rogue cell behavior (e.g., cancer)?

2.2 Biological Evolution and Cancer

In 1977, Richard Peto identified an interesting paradox in cancer incidence rates across species. Now known as "Peto's paradox," he found that larger and longer-lived species do not display increased cancer rates proportional to their larger number of cell divisions [9]. The current leading hypothesis explains this pattern via natural selection for differential cancer prevention in these larger, longer-lived species [10].

According to evolutionary biologist Leonard Nunney [11], the complexity of genetic control over unregulated cell growth should depend on a tissue's size and its pattern of proliferation. He showed mathematically that the levels of somatic mutations in small and large animals are so different that there is not a mechanism to prevent cancer in one that would be evolutionarily stable in the other. Furthermore, within species, lineage selection suppresses cancers causing the greatest loss of fitness (which may be tissue specific) [11]. Because only the germline genome is passed on to an individual's offspring, deleterious mutations in it often cause a greater reduction in fitness than somatic cell mutations. In fact, the somatic mutation rate in both mice and humans is almost two orders of magnitude higher than the germline mutation rate [12].

Based on this biological evidence, we believe that natural selection pressures caused the evolution of cancer suppression mechanisms in natural multicellular organism. Next, we sought to replicate these pressures in Avida 4's evolutionary computation environment.

3 Approach

Based on this existing literature, the goals of this project were as follows:

1. Update Avida 4 to allow single cells contained within multicells to replicate over one another.
2. Run the program on MSU's HPC.
3. Analyze the results for evidence of multicells evolving restraints on cellular reproduction that reduce or eliminate the chances of a single cell "going rogue" and wiping out the multicell.

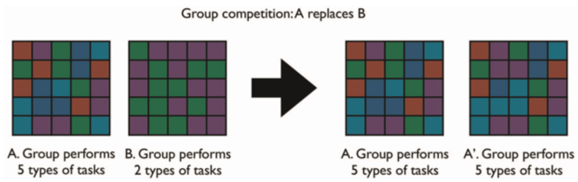

Figure 1: By performing 5 types of tasks instead of 2, group A is able to replicate over group B [1]

3.1 Avida 4 Digital Evolution System

In the Avida 4 system, digital organisms compete for space in the environment. Each organism is a fully functional computer program, set up as a circular list of instructions, and a virtual CPU that executes the instructions. These instructions make up the digital organism's genome and determine its behavior [1].

In order to self-replicate, the genome includes a sequence of instructions to create an offspring. This emulates asexual reproduction in the natural world. When a digital organism reproduces, a neighboring location is selected from the environment to place the offspring, and (if overwriting is allowed) any previous inhabitant of the target location is replaced (killed and overwritten). If overwriting is not allowed and the location is already occupied, the copy operation is canceled. Copy mutations (substitutions, insertions, and deletions) may occur during the replication process, leading to offspring that are genetically distinct from their parent [1].

The genome of a digital organism can include a variety of different instructions from the Avida instruction set. The instruction set is set up so that any combination of instructions is a syntactically correct program (though it may not perform any meaningful computation). It includes several instructions to facilitate distributed problem solving and transmit epigenetic information to offspring [1].

The rate at which an organism's virtual CPU executes its instructions is determined by the organism's metabolic rate. An organism can perform a bitwise Boolean logic operation on 32-bit integers to consume resources that increase its metabolic rate. Performing NOT or NAND doubles the organism's metabolic rate, AND or ORNOT triples the metabolic rate, and OR quadruples it. An organism may only receive a reward for performing one task. For example, a digital organism could not be rewarded for performing NAND and subsequently be rewarded for performing ORNOT [1].

Multicells are groups of digital organisms (cells). They compete with other groups and replicate via tournament selection. The inter-tournament period length determines how often this takes place. Within each tournament, the group that performs the greatest variety of tasks is replicated to the next group-generation. However, a group may only replicate if it has accrued resources equal to or greater than the group replication threshold value [1].

3.2 Avida 4 Configuration

Avida 4 uses two large C++ library sets: EALib and Boost. EALib is a series of C++ libraries for building evolutionary algorithms [13]. It includes Boost as a dependency. Boost is a set of open-source C++ libraries for general development [14]. It is notoriously difficult to install (C. Ofria, personal communication, 2019), and doing so caused significant development delays during the project.

Once Avida 4 was installed, the initial configuration was set to that used in *The Effect of Conflicting Pressures on the Evolution of Division of Labor* [1]. Each run contained 1000 5-by-5 multicell grids initially seeded with one cell each. The group replication threshold was set to 500 in the first run. For both the control group and the treatment group, 300 runs were conducted, each lasting one million updates.

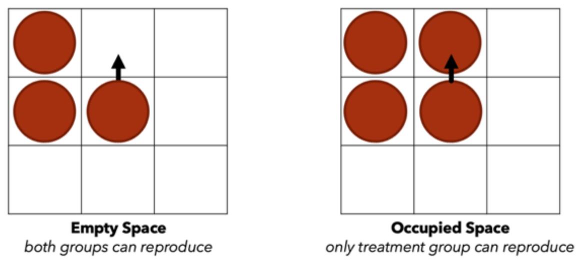

Figure 2: Tissue accretion (cell reproduction) in the control and treatment groups. Only treatment group cells may overwrite living cells.

In the control group, individual cells within a multicell could not overwrite existing cells. When they execute their copy instructions to reproduce, the program checks if the adjacent square they are facing is occupied by a living cell. If it is, the instruction is not completed. This programmatically prevents this type of rogue cell behavior.

In the treatment group, overwriting neighboring cells is permitted. This allows cells to replicate over a neighboring cell, even if it is still alive. Overwriting existing cells is a type of rogue cell behavior, since it increases the cell's fitness without increasing the group's fitness.

The first run revealed slight difference between the control and treatment groups, but not enough to understand what was going on. Therefore, in the second run, the group replication threshold was increased to 1000 to increase the pressure on multicells to accrue more cells within their grids, and the runs lasted two million updates to allow the time to do so. The results were startling, showing data that could not be possible if the Avida 4 program was working correctly.

A bug was found in the Avida 4 code base that was affecting the random number generators used by the digital organisms. Luckily, the bug had been injected after the most recently published paper, so no publications had to be redacted (H. Goldsby, personal communication, 2019). The data generated for this project was discarded, and a third run was constructed with the corrected Avida 4 program.

In the third and final run, the group replication threshold was set to 600, because 1000 in run two appeared to have been too high for the groups to evolve multicellularity. The runs still lasted two million updates to give the multicells time to evolve cancer-like behavior and, hopefully, restraints on it.

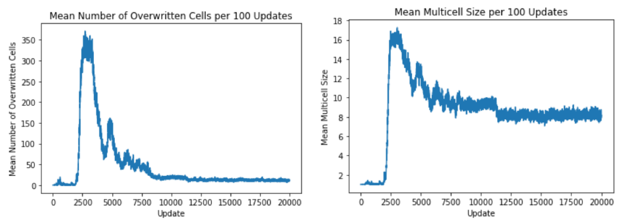

Figure 3: Cancer-like rogue cell behavior is shown in the mean number of overwritten cells compared to the mean multicell size. The number of overwritten cells peaks at over 350 before the multicells are able to evolve restraints on rogue cell behavior. The number of overwrites then drops dramatically.

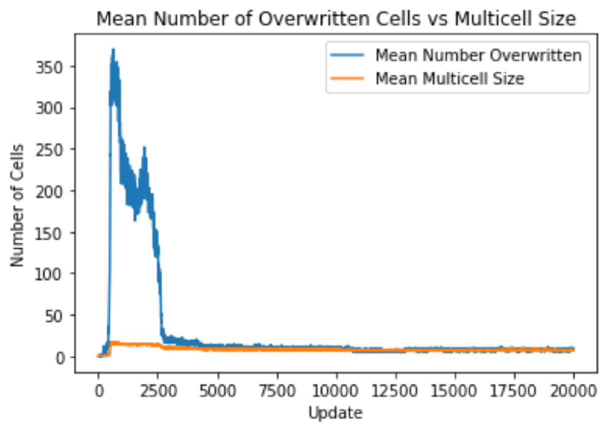

Figure 4: Compared on the same graph, the same cancer-like behavior can be seen in another run, evident in the spike in overwrite numbers. The mean multicell size is does not change enough to explain the rapid decline in the number of overwrites.

4 Results

In the third run, there was not a statistically significant difference in how often runs evolved multicellularity between the treatment and control groups . Both evolved multicellularity in about 3 percent of the runs (9 out of 300 in the control group and 10 out of 300 in the treatment group). Cancer-like behavior was observed in eight of the ten treatment group runs that evolved multicellularity. Figure 3 depicts one such run.

In Figure 4, the multicell size only drops from 16 to 8, while the number of overwritten cells drops from over 350 to about 20. This dramatic decrease is evidence that the multicells evolved some form of restraint on multicellularity that inhibits rogue cell behavior. How this restraint is achieved is a question that should be addressed in future work.

Two of the ten runs that evolved multicellularity did not evolve strong evidence of cancer-like behavior. They both had a smaller mean multicell size and evolved multicellularity later than the other multicellular runs. For example, Figure 5 shows a run in which the mean number of overwrites peaked around 30, and the mean multicell size stayed in line with the number of overwrites.

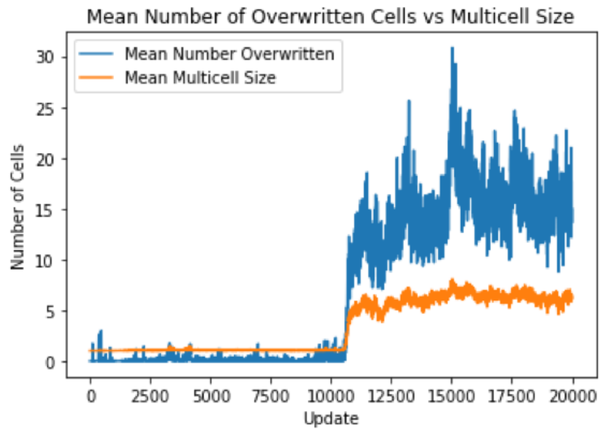

Figure 5: Two runs evolved multicellularity but did not show strong evidence of cancer-like behavior as evidence by a low mean number of cell overwrites.

5 Conclusion and Future Work

Avida 4 is accepted by both the biological and evolutionary computation communities as a representation of evolution [1]. It has been successfully used to study the major evolutionary transition from single- to multicellularity thanks to the ability it grants researchers to explicitly define and measure the digital organisms, environment, and evolutionary pressures such as the group reproduction threshold and tournament size [1]. This project shows promising evidence of the evolution restraints on multicellularity in 2.67% of the treatment group runs. More and longer runs are necessary, however, to get enough data to be statistically significant.

Increased data collection and an improved data analysis pipeline is also necessary to distinguish advantageous from disadvantageous overwriting. Questions to investigate include:

- If a cell is surrounded, does it stop replicating?
- How much of the observed behavior is limits from resources vs. limits on multicellularity?
- Why do some runs evolve cancer-like behavior and others do not?
- Is there a difference in restraint shown between germ and soma cell overwrites?
- What restraint techniques are used?

During the experiment, two unexpected findings arose that did not appear to directly relate to the evolution of restraints on multicellularity, and thus were not explored in depth. First, the control group did not evolve higher levels of multicellularity than the treatment group, despite not having to contend with potential rogue cell behavior. When the control group's group replication threshold was 500, about 10% of runs evolved multicellularity. In the third run, however, when the group replication threshold was 600, only 3% of runs evolved multicellularity. We expected the increased group replication threshold to increase, not decrease, the percent of runs that evolve multicellularity. Future work should manipulate the evolutionary pressures and lineage tracking to examine why this is not the case, and why programmatically prohibiting rogue cell behavior does not provide any apparent boost in the ability to evolve multicellularity (under the current conditions).

Second, the Avida 4 program began using more than twice as much memory per run when additions were made to track cell births and overwrites. Future work should investigate a potential memory leak. If it does not appear to be an error in the code, it is possible that the cells are utilizing communication instructions more frequently as part of their reproduction management strategy. Data collection should be added that tracks message passing between cells to examine how they are being used. Ideally, in either case, the code should be updated in a way that reduces the memory required for each run.

Studying evolution in action using digital organisms gives us the ability to understand the natural selection pressures that cause the evolution of cancer suppression mechanisms. This work is relevant to the fields of biological evolution and evolutionary computation, as well as veterinary and human medicine. The promising results from this project are therefore being built upon with research that hopes to further understand the evolutionary pressures at work and the evolved restraints on multicellularity.

References

[1] H. J. Goldsby, D. B. Knoester, B. Kerr, and C. Ofria, "The effect of conflicting pressures on the evolution of division of labor.," *PLoS ONE*, vol. 9, no. 8, p. e102713, 2014.

[2] S. Okasha, "Why won't the group selection controversy go away?," *The British Journal for the Philosophy of Science*, vol. 52, no. 1, pp. 25–50, 2001.

[3] C. J. Goodnight and L. Stevens, "Experimental studies of group selection: What do they tell US about group selection in nature?," *The American Naturalist*, vol. 150, pp. s59–s79, 1997.

[4] D. C. Queller, "Cooperators since life began," *The Quarterly Review of Biology*, vol. 72, no. 2, pp. 184–188, 1997.

[5] R. Axelrod, D. E. Axelrod, and K. J. Pienta, "Evolution of cooperation among tumor cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 36, pp. 13474–13479, 2006.

[6] C. Ofria and C. O. Wilke, "Avida: A software platform for research in computational evolutionary biology," *Artificial Life*, vol. 10, no. 2, pp. 191–229, 2004.

[7] H. J. Goldsby, A. Dornhaus, B. Kerr, and C. Ofria, "Task-switching costs promote the evolution of division of labor and shifts in individuality," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 34, pp. 13686–13691, 2012.

[8] H. J. Goldsby, D. B. Knoester, C. Ofria, and B. Kerr, "The evolutionary origin of somatic cells under the dirty work hypothesis," *PLoS Biology*, vol. 12, 2014.

[9] R. Peto, “Epidemiology, multistage models, and short-term mutagenicity tests,” *International Journal of Epidemiology*, vol. 45, no. 3, pp. 621–637, 2016.

[10] R. Noble, O. Kaltz, and M. E. Hochberg, “Peto’s paradox and human cancers,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1673, p. 20150104, 2015.

[11] L. Nunney, “Lineage selection and the evolution of multistage carcinogenesis,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 266, no. 1418, pp. 493–498, 1999.

[12] B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg, “Differences between germline and somatic mutation rates in humans and mice,” *Nature Communications*, vol. 8, no. 1, pp. 1–8, 2017.

[13] D. Knoester, “dknoester/ealib,” 2019. original-date: 2012-06-04T15:17:36Z.

[14] R. Rivera, “Boost c++ libraries,” 2019.

Chapter 6

Submission 6: Rodriguez

On the Study of Digital Evolution: Definition, Process, and Application

by Santiago Rodriguez Papa

I have been working at the Michigan State University Digital Evolution Laboratory for almost two years now. I was initially hired as a software development assistant—basically, I was to aid graduate students in their research by helping them design, develop, and debug their code. Some might find this counterintuitive, but while I had relatively good experience with C++, Javascript, Python—the three languages used in the laboratory—and a strong background in programming for my age, I had little-to-none experience with digital evolution. This is not to say I was unfamiliar with the general concept, but rather that I was not knowledgeable in its terminology, main theories, or real-world applications. Over time, I became much more knowledgeable and involved in the field. This report will focus on defining the field, explaining the process by which we study evolution, and finally giving some real-world applications.

Most people have a relatively good idea of what evolution entails. However, the general public often misunderstands certain key concepts that are invaluable for comprehending digital evolution in particular and evolution in general. Evolution is, at its core, an intergenerational (read, long-term) process by which species adapt to their environment. (MacColl, 2013). This process is fueled by *natural selection*, a mechanism by which the genetic makeup of a population is changed within a generation (MacColl, 2013). In order for natural selection to occur there must be selective pressure—that is, an outside influence that steers the population in a certain direction.

The most well-known example for selective pressure is the story of the short-necked giraffes. The tale begins by explaining that, once upon a time, giraffes used to have normal-length necks. They would roam the steppe eating leaves from ground-level bushes and short-trunk trees. One day, a fire wiped out most of this dwarf fauna. As such, the giraffe population soon ran onto the problem of not having enough food to sustain itself. It was observed that after some generations, the dwindling population began to grow in height. Soon, the giraffes grew taller on average and the whole population got access to a new source of food: tall trees.

While of course this not how giraffes originally evolved their tall necks, this example serves as a clarification to some common misunderstandings. At first, the giraffes' average height resided somewhere between the low and high trees, and the height distribution could be depicted by a bell curve. After the fire, however, the lower end of the bell curve was unable to secure food and thus reproduce, so their "short genes" were passed on less often; we say these short giraffes had *lower fitness.* On the other hand, the giraffes on the higher end of the normal distribution did have access to a reliable source of food, so their genes *were* passed on; these individuals had *higher fitness.* Since before the fire, the population height had a normal spread, we can assume that the mutations that affect a specimen's height are also normal. However, the selective pressure made it so that only the high fitness individuals could survive, and as such, the spread skewed right.

While it might seem counterproductive to define so many ecological terms in a report about "digital" evolution, I promise it is not in vain. The field borrowed the majority of its terms from biology, so it is imperative to be familiar with them. In fact, ecological and digital evolution are so closely related that, in their 1999 paper, Adami et al defined digital evolution to be the use of a computational medium in order to study evolution in action. It could thus be argued that both fields study the same thing from different perspectives; ecology focuses on biological records, and digital evolution on computational ones.

In a later paper, the same authors also argue that, while traditional methods for its study focus on an *a posteriori*, indirect analysis of fossil records and sedimentary patterns, the use of a simulated virtual environment allows for a faster and more direct approach centered on testing generalizations about living systems (Lensky et al, 1999).

Indeed, the primary use of digital evolution is to study real-world evolution at a scale not possible in real life. Since the speed at which the virtual organisms reproduce—and thus mutate and evolve—is only limited by how fast the simulation can run, we can study evolutionary effects that we otherwise would be unable to.

When asked about what I do, I often joke that I work with *Digimon*. Honestly, this is not all too inaccurate—my job is probably the closest we will ever get to that franchise. After all, the study of digital evolution can only be accomplished if there are digital organisms that evolve.[1] These digital organisms are also known as *artificial life*[2] (Lenski, Ofria, Pennock; 2003).

In our studies at the laboratory, we define digital organisms to be programs that execute basic machine language instructions. Most of these instructions are commonly found in every CPU—think ADD, POP, JMP—others are only relevant to our field of study—things like REPRODUCE or GET RESOURCES. As you can see, these programs are very similar to real-life organisms: they have genetic code that allows them to interact with their environment, reproduce, and die—this is all you need to fulfill the definition of life! In a way, you could say these beings *are* alive, just in a different way than real-life organisms.

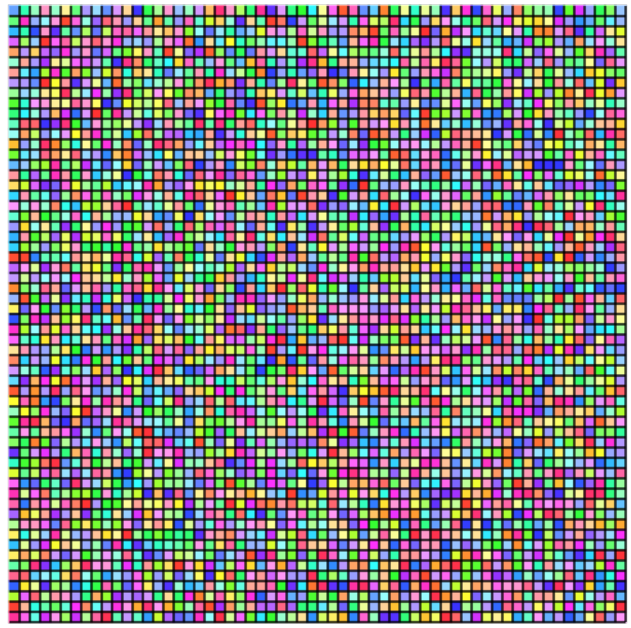

Figure 1. A randomly initialized digital world.

[1] Without organisms nothing can ever evolve, as evolution involves change within a population.
[2] Not to be confused with Artificial Intelligence.

To study evolution, we utilize a finite-sized world. Initially, this world is fully populated by cells with randomly generated code [Figure 1].

In evolutionary ecology, there are different moments in time considered to be key in the evolution of life as we know it—these are known as *major transitions in individuality* (Stuart et al, 2015). Commonly, they include things such as the grouping of aminoacids into genes, the cooperation of genes to form genomes, the transition from prokaryotic to eukaryotic cells, and the cooperation of unicellular organisms into multicellular ones. With Matthew Andres Moreno—current doctoral student and my direct supervisor—we have been studying this last transition.

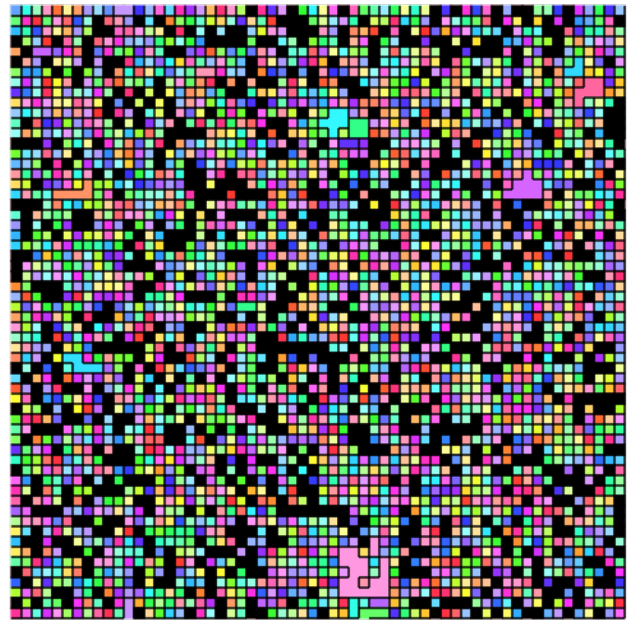

Figure 2. The same world as Figure 1 after 5000 updates. Some multicellular organisms can be observed, bottom-center and top-right. Black squares are cells that experienced apoptosis.

In order to do observe the evolving cells in a user-friendly manner, we visualize the different *filial groups* with different colors. As such, initially, the organisms are single-celled individuals. This fact can be appreciated in Figure 1 since no square shares colors with its neighbors.

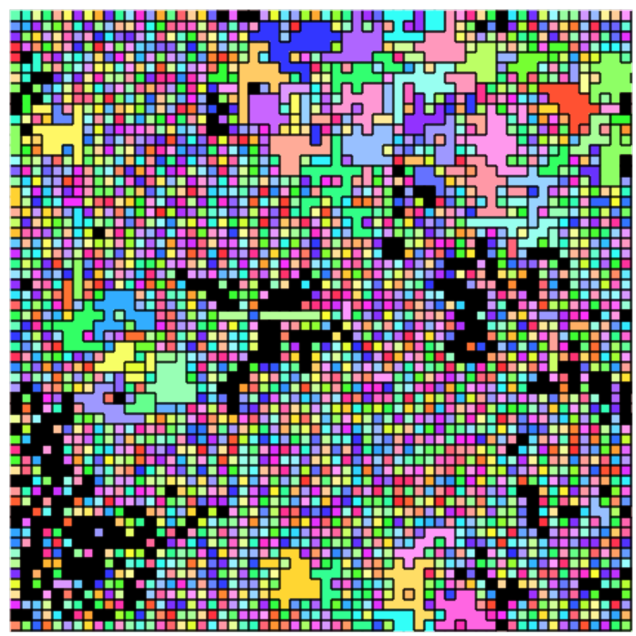

Figure 3. The same world as Figure 1 after 10000 updates. Many multicellular groups can now be observed, particularly top-right and bottom-center. The fact these groups are separate shows the transition to multicellular life occurred multiple times.

On each update, one instruction of code of each individual cell is run. If REPRODUCE is reached, the cell chooses a random neighbor to overwrite. Cells can communicate with each other through the sharing of *resource* (Moreno and Ofria, 2019). Over time, some cells will die, while others will reproduce and cooperate—when this happens, we are observing the transition from unicellular to multicellular organisms. This can be appreciated in figures 2 and 3. This transition happens repeatedly and in different parts of the world, akin to real life (Eiben and Smith, 2013).

We have answered both *how* and *why* we want to study digital evolution. We have also discussed its main application: digital evolution is used to study evolution in action. However, we have yet to discuss some of its secondary applications, which are more relevant to the field of Engineering.

The main secondary application of digital evolution is the development of algorithms. Every engineer is familiar with the concept of an algorithm: an ordered set of instructions that are executed to fulfill a certain task. However, the development of these algorithms varies. While most choose to either use previously engineered algorithms or to design their own, there is a growing concept in the field of Computer Science called evolutionary algorithms.

An evolutionary algorithm is defined to be the application of natural selection to develop some desired emergent behavior (Eiben and Smith, 2015). The software engineer must have a specific goal in mind (for example, to approximate a given set of points on the plane). After defining the goal, the programmer develops a *fitness function* to test the population against (in this case, it could be the mean squared error between generated data and known data). Finally, a population is randomly generated and the simulation is started. After a set number of updates has elapsed, the specimens' finesses are measured and a group of them is chosen with a defined *selection scheme*. This is repeated until the fitness of the population reaches the desired level. At the end, the population is composed of individual algorithms that can effectively solve the original goal.

This idea of competing algorithms against each other has many uses in the world of Computer Science. To those familiar with the topic, the previous description might seem strangely similar to the method in which Neural Networks are generated. Indeed, the fields of Digital Evolution and Neural Networking are closely related. For example, evolutionary algorithms have been successfully utilized to construct recurrent neural networks (Angeline et al, 1994). In fact, I am currently conducting independent research on the use of digital evolution to generate highly accurate neural networks in a shorter amount of time.

In conclusion, the field of digital evolution is wildly diverse and unique. This intersection of Biology, Ecology, and Computer Science might seem strange at first glance, but a closer look reveals a growing group of researchers that are revolutionizing how we understand evolution, computation, and ecology. I am extremely thankful to be part of it.

References

Adami, Ofria, & Collier. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences,*, 4463-4468.

Angeline, Saunders, & Pollack. (1994). An evolutionary algorithm that constructs recurrent neural networks. *IEEE Transactions on Neural Networks*, 54-65.

Eiben, & Smith. (2013). Multicellularity arose several times in the evolution of eukaryotes. *BioEssays*, 339-347.

Eiben, & Smith. (2015). What is an evolutionary algorithm? *Introduction to Evolutionary Computing*, 25-48.

Fortuna, Zaman, Wagner, & Ofria. (2013). Evolving Digital Ecological Networks. *PLoS Computational Biology*.

Lenski, Ofria, & Collier. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 661-664.

Lenski, Ofria, & Pennock. (2003). The evolutionary origin of complex features. *Nature*, 139-144.

MacColl. (2013). The ecological causes of evolution. *Trends in Ecology & Evolution*, 514-522.

Moreno, & Ofria. (2008). Understanding Fraternal Transitions in Individuality.

Moreno, & Ofria. (2019). Toward open-ended fraternal transitions in individuality. *Artificial life*, 117-133.

Moreno, Papa, R., & Ofria. (2021). Conduit: A C++ Library for Best-Effort High Performance Computing.

West, Fisher, Gardner, & Kiers. (2015). Major evolutionary transitions in individuality. *Proceedings of the National Academy of Sciences*, 10112-10119.

Wilke, Wang, Ofria, & Lenski. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 331-333.

Chapter 7

Submission 7: Gutai

The role of Evolutionary Biology in Artificial Intelligence and Alife

Introduction

Since its advent, Artificial intelligence has continually taken influence from Biology. However, key biological processes remain underappreciated for their potential in solving complex problems, due to a lack of collaboration between the two very distinct fields. For both biological and artificial life, finding the best solution involves navigating a highly dimensional, and often dynamic, fitness landscape. Hill-climbing techniques can result in stagnation at local optima, and variations of random search can be equally as ineffective at locating high performing regions. But nature continues to complexify, lending inspiration in tackling the most ambitious goal in AI: the creation of Artificial General Intelligence.

Combining Learning and Evolution

ANNs, which are commonplace in modern AI, superficially model the interactions between real neurons in the brain where the strengths of synaptic connections encode knowledge and experience [1]. In AI, these connections, or weights, are adjusted based on their error from the desired performance [2], working similarly to how intrinsic rewards systems modify synaptic weights in animals [3]. The development of ANNs caused a lot of initial excitement but it was soon realised to have limited success in producing the complexity of human cognition. Brains have been optimised over millions of years of evolution to have a stereotyped connectional architecture which forms the basis for learning. The product of this evolutionary optimisation is the diverse range of abilities animals are born with such as walking at birth or basic capabilities we take for granted such as objectness. Whereas conventional ANNs rely on just learning, resulting in a number of deficiencies [4].

Genetic Algorithms (GAs) work by simulating survival of the fittest. Populations of bit strings, each representing a solution, evolve and occasionally mutate and only the fittest solutions can mate and propagate to the next generation [5]. Mating works similarly to crossing-over in meiosis; a random point along the string is selected and bits are swapped creating two new offspring. Closely adjacent sequences are less likely to be separated by crossing over, like in genetic linkage, so tend to be synergistic and form partial solutions, like genes. Recombination of good partial solutions enables the search to 'jump' between optima focusing on the most promising parts of the search space, while purging poor partial solutions which would otherwise accumulate [6]. Neuroevolution applies GAs to adjust the weights of ANNs in place of backpropagation. A GA evolves a population of bit strings which represents the connection weight matrix of an ANN. The fitness of the bit strings are determined by the performance of their resulting ANNs at a given task. Neuroevolution emerged in the 90s [7] and GAs are now argued to be a competitive alternative to backpropagation [8].

Just as pure learning has its deficiencies, so does pure evolution. Hill climbing techniques such as backpropagation enable a local optimum to be reached but the global optimum may never be discovered. In contrast, pure evolution is more likely to locate the region of a global optimum but won't necessarily locate its peak because it does not directly follow a local fitness gradient. As such, these approaches are complementary, just as in nature. Evolution shapes the fundamental innate behaviour based on the experiences of ancestors, and learning optimises it so it can best be used – as the precise conditions and events experienced by each individual in a lineage will differ subtly. Coupling evolutionary algorithms with deep learning strengthens this analogy with nature and exploits the best qualities of each approach. This may be done by using GAs to evolve the general structure of the ANN and training with backpropagation to fine tune the parameters.

Evo Devo and Indirect Encoding

For common AI, be it deep learning or neuroevolution, learning occurs directly on the parameters of the ANN. In contrast, the human genome holds only 1.5GB of information; this is not enough to direct encode the trillions of synaptic weights in the brain, let alone the entire organism. However, the genome does not explicitly encode the body map of the organism, it encodes its development. The result is an organism that has modulated repetition with variation. For example, most multicellular organisms display some form of symmetry and numerous features which are derived from a single innovation, e.g., insect appendages: wings, antennae and mandible, are all variations of the leg appendages. This property enables coordinated evolution of structure and behaviour.

HyperNEAT [9] incorporates indirect encoding into Neuroevolution to evolve perfectly coordinated walking gaits of simulated creatures. Like in classic Neuroevolution, Populations of bit strings evolve, mutate and mate. These bitstrings do not encode the weights of the functional ANN as in classical Neuroevolution, but they encode the weights for a Compositional Pattern Producing Network. The output of this network is a geometric pattern with repeating motifs, and it is the coordinates of this pattern which encodes the weight of the functional ANN. The resulting ANN is able exploit regularities in a problem; for example, HyperNEAT can automate the evolution of highly coordinated four-legged gaits [10], which has only previously been possible with manual decomposition of the task.

Group selection and a division of labour

From a genetic to a societal level, group selection has repeatedly resulted in major evolutionary transitions. For example, the eukaryotic cell arose from a proto-eukaryote's engulfment of and subsequent mutualism with a prokaryote which became the mitochondria. The reproduction of the mitochondria eventually depended on the reproduction of the eukaryotic cell and *vice versa*

[11]. Due to their aligned reproductive fates, there is a selection for cooperation and this manifests in a division of labour – the mitochondria provide optimal conditions for aerobic respiration and the host capitalises on this while providing the energy and resources required. This example of cellular compartmentalisation enables much greater efficiency within the cell and facilitates further beneficial variation [12].

At the genetic level, information is encoded in linear DNA sequences composed of four bases. Although having a simple linear structure, sub-sequences encode higher-level functional elements such as genes, and these are further grouped together into chromosomes that ultimately make up the final genome of an organism. Each component in this hierarchy has a distinct role in synthesis and the proliferation of the others and so success requires cooperation between them [13]. A similar phenomenon also occurs during the evolution of GAs where bitstrings are used to encode the genome of a potential solution. Crossing over allows for partial solutions to become localised within a string such they can propagate effectively. These sub-sequences can then be further combined at larger length scales establishing a hierarchy of modularity even within this simple system. By viewing a complex task as a collection of smaller interacting subtasks, it is possible to tackle larger problems in smaller chunks, facilitating search in even highly deceptive spaces. Because of this, genetic algorithms have been found to better tackle tasks with complex fitness landscapes compared to ANN-based deep learning approaches.

At higher levels of organisation, social groupings are common between organisms (e.g., parental care, herds and flocks). There are many benefits to group living. An extreme example of this are the eusocial insects whose colonies form what are termed a 'superorganism'. These colonies are composed of distinct castes – such as queens, workers, and soldiers – which are morphologically and behaviourally specialised to their task. In AI, a problem can be tackled using multiple cooperating agents if fitness can be measured by the sum of their behaviours.

For example, OpenAI created a human-relevant example of how this works using a game of hide and seek [14]. Pairs of seekers played against pairs of hiders where each is controlled by its own ANN. The hiders worked together to build a shared wall around themselves before the seekers finished counting. Once the seekers discovered they could use ramps to breach the wall, the hiders exhibited a division of labour where one began constructing the wall while the other confiscated the ramps – a strategy that would have be impossible by either alone due to countdown time limit.

The hierarchy of modularity in current artificial systems is modest compared to that found in biology. In nature, group selection occurs on numerous levels simultaneously, establishing an extensive hierarchy of task decomposition which scales with complexity covering genomes, endosymbiosis, cell compartmentalisation, multicellularity [15], tissues and organs, organisms, sex [16], social groups, and inter-specific mutualisms [17] to name a few. In effect, a complete ecosystem also exhibits modularity through niche differentiation, which operates to reduce competition.

Complexification

A few billion years ago, life began simply – morphologically, behaviourally and genetically in a simpler environment and with little competition. From then on, the elaboration of genomes enabled new behaviours and structures. This adaptive radiation of life has created multiple kinds of organisms with different metabolic processes occupying different niches. As these organisms interact with their environment, they cause changes opening up new niches for further diversity to arise. This leads to a continually growing web of interactions with multiple trophic levels. Conventional evolutionary computing infamously lacks this open-ended quality, usually halting as it stagnates at a final solution. Often, evolutionary computing techniques involve a single static selection pressure based on the performance of the task of interest. Yet the complex behaviours we are interested in recreating in AI such as language are the product of gradual complexification, rather than an explicit selection pressure for language in isolation.

Novelty search has shed light on the shortcomings of this conventional approach [18]. Behaviours are selected based on their uniqueness rather than their performance, and once the search for the simplest behaviours is exhausted, more complex behaviours are generated, promoting diversification and complexification. The selection for novelty occurs in adaptive radiation, where species differentiate and specialise into niches to reduce competition. In a biped locomotive task, novelty search significantly outperforms fitness-based approach despite being agnostic about what behaviour is adaptive for walking. This study highlights the deficiency of using a single fixed selection pressure based on task performance – by ignoring the objective, it's possible to explore deceptive domains associated with the task fitness.

The evolution of the brain in particular, from smaller primitive brains to complex ones, inspired the development of ANN topological complexification in Neuroevolution of Augmented Topologies (NEAT) [19]. ANNs evolve from minimal architecture and, in addition to adjusting the weights, mutations affect the topology of the network by adding new nodes and connections. Therefore, optimisation occurs simultaneously with complexification, mirroring how our brains evolved from primitive smaller ones, or how mature brains develop from limited-capacity infant brains [20]. NEAT significantly outperforms neuroevolved ANNs of fixed topology in pole balancing, a standard benchmark learning task. In addition, the evolution of topology eliminates reliance on humans to choose the most appropriate topology prior to evolution.

A major drive for complexification is antagonistic coevolution. Predator and prey, for instance, perpetually create new selection pressures for their counterparts. A famous example of this is the arms-race between bats and eared moths [21]. The ability of bats to locate moths using sonar clicks drove the evolution of sonar detection in moths. This was followed by bats altering the intensity and frequency of their sonar click for stealth. Moths then evolved their own sonar clicks to 'jam' the frequency of the bats. Such arms-races are never-ending, a phenomenon known as 'the Red Queen Principle' [22], and result in individuals constantly adapting to keep up to new innovations. The same phenomenon occurs with artificial systems, with more sophisticated behaviours evolving and more swiftly [23].

A factor often overlooked in robotics which limits complexification is the evolution of morphology, i.e., there are a limited number of behaviours possible with a single fixed morphology. Behavioural adaptations often depend on physical adaptations. Simultaneous evolution of morphology and behaviour [24] opens up a vast possibility of phenotypes to be explored compared to using a single fixed morphology [25]. One way to capture this feature in ANNs is by employing Compositional Pattern Producing Networks which enable the structure of an ANN to dynamically evolve and explore potentially useful topologies for a task at hand [26].

Complexification occurs not only with evolution but also due to learning within the lifetime of an individual. Humans and other animals learn most efficiently if they begin with simpler tasks and gradually expand their knowledge and skills [27]. For example, musical students begin practice at grade one and slowly progress before attempting a grade eight piece. This is so necessary in producing complex cognition that it is effectively 'built in' to cognitive development: the initial cognitive limitations and slow development of some animals, particularly humans, is thought to be adaptive to facilitate incremental learning during early life, rather than just being an artefact of incomplete development [28]. For example, infants are only able to focus on objects ~10 cm away, allowing them to learn from simpler sensory information without the complication of size inconsistency and distance. This form of gradual learning can be applied to AI too: two pure-learning ANNs which play Go against one another both begin at novice level and gradually become more skilled as they play, learning better strategies and faster than if they trained against master opponents to begin with. A major benefit of self-play is that it eliminates the need to obtain human game data across a range of skill levels or the need to spend time playing against humans in real-time, which is slow and limited in scope.

It is likely that the combination of multiple kinds of complexification is necessary for open-ended evolution [29]. For example, implementing just competitive coevolution alone can cause stagnation due the loss of the Red Queen gradient, where one species discovers a strategy so strong that it cannot be beat by an opponent. Whereas, when paired with topological complexification, the space of possible behaviours continually broadens, helping to establish an effective open-ended arm-races [30]. Complexification is particularly effective using HyperNEAT because, as the problem is solved geometrically, the solutions are scalable. Gauci and Stanley found that HyperNEAT-evolved ANNs that first learn to play the board game Go on a 5 × 5 board could subsequently learn faster and discover better strategies when then faced with a larger 7 × 7 board [31].

By starting small, task performance can be optimised in a simple fitness landscape, and then by gradually complexifying the AI, the solution enters larger search spaces in more promising domains than it would otherwise have starting there. Plus, these approaches eliminate reliance on manually designing progressively more complex training tasks, as is necessary in conventional 'incremental evolution'.

Conclusion

The prospect of building AGI depends upon our ability to create systems which can work in an autonomous, innovative, and open-ended way – much like biology. Here I have discussed how evolutionary techniques, provide an attractive solution to achieving these qualities, inspired by the origin of animal intelligence. Key biological concepts continue to inspire advances in AI, suggesting that the pursuit of AGI would benefit immensely from the active involvement of Biologists in the field.

Chapter 8

Submission 8: Tanaka

100 years of Čapek's Robots

ALife 2021 - Student Essay Competition

Fabio Henrique Kiyoiti dos Santos Tanaka

In 1921, the English language was introduced to the word ”robot” for the first time when Czech author, Karel Čapek, released his theater play R.U.R. (Rossumovi Univerzální Roboti). Today, 100 years after it has been published, its influences can still be perceived in science-fiction literature and in the vision of the future regarding robots.

In the play, robots are mass-produced to handle manual labor. While some organizations fight for these artificial beings' rights, the robots themselves do not show any complaints or wants of their own in the beginning. As time passes, factories start producing more advanced robots, and their use shapes the world economy; not before long, they handle humanity's every need.

In the play's final acts, these more advanced robots start questioning why they should serve humankind and, because of that, they begin an uprising. After eliminating all but one human and taking control of Earth, the only thing the robots lack is the ability to reproduce, a secret that was lost during the extinction of humanity. The play ends when two of the robots, when faced with the death of each other, start to demonstrate affection between themselves and the last human says that this may be the secret for reproduction and that they would serve as the Adam and Eve of this new civilization. This event leads to the robots inheriting Earth as the successor of humanity.

It is important to note that, contrary to the current concept of the word, the robots in the play are less of a mechanical machine and more of an artificial biological creature. Factories for bones, nerves, arteries, and intestines are described in the story, and the process of assembly is compared to the one of an automobile. Čapek's robots are living biological beings, but they are still assembled instead of grown or born.

While this narrative was the first document to introduce the word "robot" to our dictionary, other automatons existed in the literature for centuries. In Greek mythology, Hephaestus, God of metallurgy and craftsmanship, constructed a giant man made of bronze to defend the island kingdom of Crete. In Jewish folklore, the Golem, an anthropomorphic being made of clay, could work as a servant or protector of its people. While there are many different versions of these tales, they both tell a story of constructs that follow humans' (or God's) instructions to complete the manual tasks they were designated to do.

Maybe, the story that this play resembles the most is Mary Shelley's "Frankenstein", the novel considered by many as the pioneer of science-fiction. While R.U.R. and Frankenstein were written more than 100 years apart, both narratives tell the story of humans trying to create an artificial life form and the flawed relationship that arises between the creator and its creation. In the end, both conflicts end with the death of the creators and the regret for trying to create life.

Despite the themes of these possible inspirations, R.U.R. acts as a cautionary tale for the blind search for progress, and if you consider the time it was written, it is not difficult to understand the reason. World War I have ended only a few years prior, and its impacts were still being felt. The idea that humankind would only benefit from technological progress was heavily contrasted by the newly developed mechanical and biological weapons used in the conflict. This post-war society may be one of the reasons for the author's skeptical attitude towards technological progress.

During an interview to the London Saturday Review, Karel Čapek wrote: ”I wished to write a comedy, partly of science, partly of truth. The odd inventor, Mr. Rossum, is a typical representative of the scientific materialism of the last century. His desire to create an artificial man — in the chemical and biological, not the mechanical sense — is inspired by a foolish and obstinate wish to prove God unnecessary and absurd. Young Rossum is the young scientist, untroubled by metaphysical ideas; scientific experiment to him is the road to industrial production. He is not concerned to prove but to manufacture... Those who think to master the industry are themselves mastered by it; Robots must be produced although they are a war industry, or rather BECAUSE they are a war industry. The product of the human brain has escaped the control of human hands. This is the comedy of science.”

Another theme discussed in the play is the status of robots in human society. During the narrative, the artificial humans are often referenced as slaves or serfs that lack the essence of a human being, a soul. This dehumanizing description is not too different from how some slave-owners would describe their slaves. It could be argued that the robots uprising in the play was influenced by the many revolts of slaves and serfs that can found in history. It is not a coincidence that the name ”robot” is based on a Czech word for ”forced labor.”

This parallel with human history can also be perceived in one of the proposed solutions to deal with the threat of robots during the play. When the characters thought the revolt had been suppressed, one solution to avoid future conflicts was to start producing ”national robots”. Each robot-producing factory would make robots with diverse colors, hair, and languages; the reasoning behind this was that the robots would be strange to each other and would not only be unable to cooperate between them but even hate the different robots. This solution was most likely inspired by the recurrent conflicts between people of distinct races, cultures, or nations throughout history.

Now, changing the focus from the influences to Karel Čapek writing to the impacts it caused on the media, its effects are evident until this day, 100 years later. The introduction of the word ”robot” to the English dictionary is the most notable. Although the original use referred to the mass-produced artificial biological humanoids, the Cambridge Dictionary defines it as ”a machine controlled by a computer that performs jobs automatically”. This word not only functions as a technical term, but its expressiveness can instigate the imagination of many people.

Another contribution of this work was its influence on stories that revolve around a robot or A.I. taking over. Arguably, this play was the first to tackle the idea of a human creation losing control and overpowering humanity on a global scale. Although the themes or delivery of this concept may differ, many works stay true to this formula to this day.

In contrast, Isaac Asimov, one of the most celebrated science-fiction writers of all time, was a fierce critic of Čapek's work. While he recognized the importance of introducing the word ”robot” to the global lexicon, he considered the play terrible. But, at the same time, it is possible to argue that R.U.R. had at least some influence on him when he coined the term ”Frankenstein complex” to describe humanity's fear of machines rebelling against their creators (alongside the evident inspiration from Mary Shelley's ”Frankenstein”). His famous laws of robotics served as a prevention system to avoid a future like the one present in Čapek plays, where human's creations revolt against their masters, from happening.

Nowadays, this ”Frankenstein complex” is still very present in society. From big box office movies like Marvel's ”Avengers: Age of Ultron” to popular games like Valve's ”Portal”, different media approached the idea of a robot becoming self-conscious and revolting against humans. Recent developments in A.I. technology are also a source o worry for many people. Even famous scientists like Stephen Hawking have voiced their concerns about the advancement of this type of technology.

Even though 100 years have passed since R.U.R. publications, the reasons for skepticism towards these advancements are arguably the same. They are often based on human history and the numerous conflicts and incidents present in it. Although this is a valid point of view, learning from the past when planning for the future is a characteristic of humanity, there is no foundation to affirm that a sufficiently advanced A.I. would act in an antagonist manner and oppose human beings.

Thinking that if an A.I. keeps evolving, it would inevitably start acting like humans, sounds arrogant because it poses humanity as the final or necessary step in evolution. While competitiveness and aggression are commonly thought of as required for survival from an evolutionary perspective, there are also many examples in nature where cooperation wields more fruitful results. It is not even assured that if an A.I. became self-aware, its goals would conflict with humans.

Of course, this is not to say that precautions should not be taken. While impractical, Asimov's laws of robotics are an example of how these measures are being considered long before A.I. became a popular subject in science. If a potential risk has to be pointed, deliberate "bad actors" using the technology is, and always has been, a concern. But the risk of misuse is much as a problem for A.I. as any other advancement.

To conclude, while Karel Čapek's "Rossumovi Univerzální Roboti" may not currently be the most well-known science fiction story, its influences persist. It not just introduced the word "robot", but many of its themes are relevant to this day. While the possibility of robots taking over the world is still up to debate, it is important to remember that R.U.R. was the first narrative of this type and laid the foundation for this discussion not only in literature but also in science.

Chapter 9

Submission 9: Gallotta

On why seeking Artificial Life

An essay by Roberto Gallotta

The field of Artificial Life is populated by many different scenarios and environments, spanning from swarm robotics (Hard ALife) to chemistry simulations (Wet ALife). This is caused mainly by the fact that the entire field is evolving, just as the subject of its study evolves. While all these scenarios have interesting properties and applications, in this essay we will reason about one of the more appealing facets of Artificial Life: the evolution of digital creatures in a simulated environment (Soft ALife). Inspired by biology, these creatures are part of a population and evolve into different species, each characterized by different behaviors when presented the same stimuli, while performing a given task that determines their fitness. One important aspect of such evolution is its desired open-endedness: the creatures should be able to evolve and show an ever-complexifying behavior, just like real life evolved creatures have shown to do in the course of history. This behavior however is closely related to the programmer-defined task. In general, we can say that the "life" component of this simulated environments is the task itself.

There is a problem here: the task that we would like to solve has to be reflected in the environment we are able to produce, but the environment cannot always faithfully reflect the task in its entirety or in its full complexity. This is an obvious compromise between what we would like the creatures to do and the limitations of the hardware these simulated environments run on. For example, while it could be possible that a creature may learn to walk with a specific gait in an incredibly rich environment where it could interact with other creatures, it would be superfluous: a simple plane with no inter-individual interactions would work just fine for the task, though it clearly lacks the requirements for open-endedness. This compromise also forces the programmer to introduce a bias in the environment. Thus, it is safe to assume that the environment design is as important as the algorithm that governs the evolution, if not more. While evolutionary algorithms may present problems due to their complexity or errors the programmer did not or could not foresee, errors in the environment lead to much more disastrous outcomes. A species may never evolve to a possible fitness level or it may not be able to exploit emergent behavior, but we would still be able to gain insight to its evolution and we would be assured that such insights are derived from an environment that reacts to and interacts with the individuals as expected. On the other hand, if the environment is not carefully crafted it could be exploited, effectively nullifying any effort to gain valuable insight from the population's evolved behavior. As an example, we could think of the aforementioned learning with a specific gait task. An improper evolutionary algorithm may preclude some walking patterns from emerging or from being kept down the individuals' lineage. Instead, an error in the environment such as a non-accurate physics simulation engine could be exploited, and the highest-fitness walking gait may turn out to be to exploit such physic glitches. This is not an as uncommon phenomenon as one may hope. The presented example is obviously an exaggeration of what could be a possible problem in an environment and, while in this example it would be quite easy to spot such error and fix it, in other cases similar errors may be very minute or non-trivial to spot. However, these would still affect the evolutionary process in such a way that the resulting population's behaviors reflect this error and render observation incorrect at best and misleading at worst. Accepting the importance of a proper environment, we can now assume that extreme care is taken in developing the best environment for the task.

Now that we assume that the environment is always the best for the task, we can turn our attention to the evolving individuals. We particularly look at their possible designs and limitations in such designs. Just as for the environments, the individuals' design complexity is limited not only by the designer's choices, but mainly by the computational constraints. The designer's choices are an obvious reflection of the task requirements, whereas the computational constraints are but a temporary hurdle that could be overcome as technology progresses in both power and availability. Both limitations, however, directly influence the observable behavior of the evolved individuals. It is not farfetched to assume that a designer consciously modifies their original designs during development due to the computational constraints. While it would be appealing to say then that those computational hurdles are already integrated in the designer's bias of their product, this is not always the case. It is in fact entirely possible that an individual's design for a specific task could work in a feasible amount of time now rather than 30 years ago. For example, an individual whose behavior is controlled by an evolvable artificial neural network would not have been implemented 30 years ago because such design was simply not as efficient as it is now, so running a large number of generations would have been unimaginable and, thusly, no study on the emergent behavior could have been conducted. This phenomenon is widespread in the machine learning community as the "hardware lottery", where one solution to a problem can be found in practice thanks to the current level of technology and available computation power and a different solution instead is less performing because the current hardware does not support it as well as the other. However, while in the machine learning field this just means that we must come up with different architectures that may just perform sub optimally, in the artificial life field this means that a different design must be implemented, and we have no assurance that we are not precluding any interesting behavior from emerging at any point of the evolution. If our interest is the study of emergent behavior, then this problem is incredibly concerning. This point, at least currently, is mainly theoretical as current computational power is on par with common tasks requirements. If instead one encounters such a problem where their technology cannot handle their design, it probably would be best to tune different parameters first, such as the size of a population or other algorithm-specific parameters. It is then important to distinguish between the types of possible individual designs. These can be divided in two main groups: the first has a fixed structure and the evolution process simply controls the parameters within it (think of a neural network where the architecture is fixed but not the weights and biases), the other provides a set of base building blocks and the evolution process controls how many of these blocks are there, their connections and their internal parameters (think of the networks evolved with algorithms similar to NEAT). Although apparently different, both possible designs produce behaviors that reflect the designer's bias. In the first case this is more apparent, as the structure is handcrafted and implemented *a priori* by the designer and any alteration to this structure results in different output behaviors. The designer's influence in the second case is slightly more subtle: one may think that since the evolution process governs the structure overall design as well as its parameters, then the designer's bias is washed away and any emergent behavior is a direct result of the environment and the algorithm that guides the evolution process. This is not the case: while on a larger scale the architecture is constantly modified and mutated, its base blocks are not and changing these base blocks would change the output behavior of the entire structure. One could argue that changing the base blocks does not necessarily mean that some solutions may never be discovered, but the reality is that we simply do not know *a priori* if that would be true or not. The same solution may be found with a more complex structure that takes much longer time and resources to evaluate. It would be interesting to experiment changing these base blocks in a domain and see what, if any, behaviors do not appear anymore during the evolution. However, both approaches to designing the individuals' architectures are viable and, depending on the task, one may work better than the other. The second approach, however, is probably the more interesting for the artificial life researcher as it more closely resembles biological evolution.

We have given motivations for carefully designing the environment and the individuals that will evolve in such environment, so we can now pay closer attention to what could be the reason to do research in the artificial life field. We first can see artificial life for the sake of artificial life: this is when the goal of our simulations is to be able to run these simulations. This is the case for games that give the user the options to control the evolution and then show the results as time goes on. A more interesting aspect is instead that of determining how certain behaviors appear in communities of individuals. This can be extremely useful for validating hypothesis on real life evolution of biological creatures. While the overarching goal of artificial life exploration is to implement biological evolution more and more accurately to solve tasks, it is also important to realize that such implementation may never be perfect. This can be cause by a multitude of reasons, but the more prominent one is, as explained in this essay in the previous sections, that computer cannot currently compete with nature's complexity. However, this is not a reason to be discouraged in seeking artificial life: any advancement in this field is important since it allows us to better understand how we evolved to become what we are now and how we are evolving to become what we will be.

Chapter 10

Submission 10: Vogrin

Do the Robots of the Future need Artificial Imagination?

By Michael Vogrin

The intriguing question "Do robots dream of electric sheep"[i] served not only as a book title for Philip K. Dick, but also further opened up the discussion of what possible "mental processes" robots could be capable of. From science fiction novels to movies and video games we find interesting explorations of this question. An elegant answer to this is hidden as a throwaway line in the video game "Borderlands 2"[ii], where a voice talking to robots roughly says: "Each robot who kills a monster, gets rewarded by getting implemented the capacity to be proud to have killed a monster.". This may lead us to believe that robots can do everything that we implement. The question of this essay is: Could we, or should we, implement artificial imagination into future robots? I want to argue two points. First, it is theoretically possible to implement something into robots that resembles human imagination. Second, this capacity for artificial imagination is neither necessary nor especially useful for robots.

Imagination is a special way of thinking. So, one might ask: Why should one think in the first place? Alfred North Whitehead gives a concise answer:

"The purpose of thinking is to let the ideas die instead of us dying."[iii]

In this quote, Whitehead uses "us" to refer to humans, but the idea expressed in his statement can be extended to other living beings, and - as I would like to argue in this essay - to robots as well. So how does thinking let ideas die, instead of us?

Imagine standing on a street, with the intention to cross the road. While doing so, it is preferable to not get hit by passing cars. To solve this problem, one might rely on technology like traffic lights, that - in conjunction with social rules - make crossing the road relatively safe. Another way is to simulate the situation, to think it through, to let it happen in your mind's eye - to *imagine*. If you can imagine yourself safely crossing the street without having to assume unusual behavior of ongoing traffic (e.g., everyone perfectly stopping just for you, or cars flying high above you to avoid a collision), you can indeed cross the street. But what if, you see yourself getting hit by a car in your imagination? Then you would refrain from crossing the street, and remembering Whitehead, you let the idea die, instead of yourself. Please note, that this may not be how we usually cross the street, because we automated this process. Nevertheless, we *can* use our imagination in this way. Could a robot do the same?

Imagine a robot on the road. Its task is to cross the road without getting hit by a car. Surely, it would gather information using its sensors such as sounds and positions of nearby objects. Usually, a robot would have a set of rules that enable it to react to information gathered by sensors. In the road-crossing scenario, it would be sensible to estimate how long it would take to cross the road. In the next step, the robot would need to check if there are cars approaching. For this, it could "look" left and right to see if there are cars in sensor range and estimate their speed. Having this data, it is possible to calculate if the time needed to cross the street is shorter than the time it would take the car to cross the path that the robot plans to take. Suppose a car is approaching at such a velocity, that it would hit the robot and destroy it. Thus, the robot correctly decides not to move. Now, did the robot *imagine* itself getting hit by a car while trying to cross the street?

I would argue that it did not imagine itself crossing the street, at least not in the way a human would. The robot produces some numerical values and reaches a conclusion, but at no point *imagines* what would happen in the case of a collision. It does not consider if the driver of the car would brake to avoid a collision. Neither does it take into consideration the possible screeching of tires or shocked gasps produced by pedestrians. In contrast, this is exactly what we humans would do. Clearly, one could argue that the robot did in fact produce an imagination, just a less vivid one, or maybe one that is not visual. As humans primarily orient themselves using vision (note the large size of the visual cortex) and robots may often be more mathematical creatures, it seems natural to assume that calculations make up a robot's mind. However, I think that only humans imagine, and robots merely conclude. In that, I see one of the main differences between how humans work in comparison to how future robots that possess the capacity for imagination will work: Humans imagine, in order to conclude, while robots conclude, in order to imagine.

Imagination is something that is indeed very human: we do it constantly and automatically. Some forms of imagination are even involuntary and have been described as "chatter in the skull"[iv] or keep us awake in the form of ruminating[v]. However, it is also tremendously useful, as it lets us live through potential scenarios so that we are more prepared for them when we must confront them in reality. This is one reason for dreaming, as it allows us to tap into necessary but possibly overwhelming chaos in the comfort of our own brain. It is a problem-solving method. Our imagination simulates scenarios, and we then conclude what outcomes are plausible, and which are not. We use what we imagine as a proxy for the future and extract the information we need from it.

So why would the imagination of a robot be any different? Because the information that we humans extract from what we imagine, is exactly the information the robot would need to construct its imagination. To build an imagination, a robot would need to first analyze the situation to a large extent. This gets clear when we go back to our example. The robot that tries to cross the street might "see" (rather: detect) that a red car is coming from the right. Estimating (rather: calculating) the speed, and judging (rather: measuring) the distance it needs to travel to cross the street, the robot concludes that it would collide with the red car. Could it solve the problem by using imagination? Could it imagine, rather than calculating and measuring everything, that a collision is probable? It could, but it would have to use exactly the same information that is detected, measured, and calculated, in order to build the imagination. This gets clear if we conceive of a robot that uses all this information, and then simulates the situation in a way that includes a visual representation of it. With just a snapshot of reality, it could simulate the road crossing scenario, and even show it to us via a screen, not unsimilar to a video game or an animated movie. But, and this is the important part, the robot would have nothing to gain from it. It makes no sense for it to look at the screen, which is equivalent to consulting its imagination, and then conclude: “Oh, I would get hit, if I cross the street now.”, because it already knows that - otherwise it would not have been able to produce the simulation.

Of course, it is possible to construct an unconventional robot that uses sensor information and then constructs possible scenarios following certain laws. It would build a visual model of the world, and within the model, there could be things like collision detection, which then are used as a proxy for what would happen in the real world. But why should we construct robots that follow this detour using a visual model, instead of just making the calculations they need to build the model directly?

After all, I conclude that it would be entirely possible to construct robots that have artificial imagination, but that it is not of any real use to them. The artificial imagination could be like human imagination in many ways: it may be built on certain laws that are acquired, it could draw from memory, it could be informed on environmental data, even be based on one's own preferences and influenced by individual biases. However, robots would not gain much from their imagination, as they - at least in principle - could have access to all the data and operations they use to build the imagination *directly*. Humans, however, function the other way round: they imagine things to extract information out of their imagination. I am aware that it might look like we, just like the robots, need a lot of information so we can build a realistic imagination. However, much of the information that we use in this process is not conscious, not readily available to us. This is *the* spectacular property of the human mind: Out of the mist that is conscious and unconscious information stored in our brain, we can condense ideas and put them into concrete action. This becomes strikingly clear when we think about problems, and -

without gathering new information - come to an answer, only by thinking. For many problems, imagination is the tool, as it lets us see with our minds eye. The construction of the image in front of our minds eye is called the imagination. It is noteworthy that the process of imagination does not need explicit consideration of natural laws, of the speed of ongoing traffic, or possible behaviors of others. On the contrary: we just imagine. We cannot really explain how we imagine, because the process is implicit, and so is the information that we weave into it. We just do it. However, clear results, concrete answers, such as to the question whether we can safely cross the street, then come into our consciousness. Ultimately, this is hinting at the astounding fact that imagination really is just a form of introspection - but one can imagine this being the topic of another essay.

Endnotes:

[i]https://en.wikipedia.org/wiki/Do_Androids_Dream_of_Electric_Sheep%3F#:~:text=Do%20Androids%20Dream%20of%20Electric%20Sheep%3F%20(retitled%20Blade%20Runner%3A,Dick%2C%20first%20published%20in%201968.
[ii] https://borderlands.fandom.com/wiki/Borderlands_2
[iii] https://www.goodreads.com/quotes/10092515-the-purpose-of-thinking-is-to-let-the-ideas-die
[iv] Tao: The Watercourse Way, Alan Watts (1975)
[v] https://journals.sagepub.com/doi/full/10.1111/j.1745-6924.2008.00088.x?casa_token=bA-8pUbDAFYAAAAA%3AZVnl58Kkrk81vyEPLUMutMax60N-_-aUgps0yQM-zFgRIvs1fJClQGGuvqlY4qJGTRi3Gomr4z0WgQ

Chapter 11

Submission 11: Hariprakash

Modern Prometheus: Can Robots ever be considered Living?

Aakriti Hariprakash

The fields of AI, ALife and robotics have been entangled since they were birthed; and to date they continuously feed off each other to thrive. This arena is on the cusp of revolution, indicated by the growing capacity of computing power every day, the recent breakthroughs in genomic data technology, and the ever growing understanding of fundamental biological mechanisms at a scale previously unknown. The real world clinical applications of synthetic biology research have the potential to dramatically alter the current model of healthcare and shift it to a more prevention based model. AI and robotics are being applied in almost every industry- from e-commerce to transportation to healthcare as well, and with the steady normalization of AI has come a host of public concerns.

As with every innovation and discovery threatening to change the status quo, people tend to look upon such change with fear and suspicion - often at the cost of scientific advancement. The common legal, ethical and social issues shared by the fields of AI, ALife, and robotics include the potential of technology misuse, the protection and privacy of data (personal as well as genomic) and the general apprehension at the idea of being superseded by technology in terms of intelligence, consciousness and longevity. This discomfort has been furthered by fiction and ignorance alike - concerns over unemployment due to automation can quickly spiral into fears of a world being taken over by robots, as in the play *Rossum's Universal Robots*. Thus, even the basic philosophical distinctions between living and non living are not only reassuring clarifications, but form the foundation that ALife and its sister fields rest on. In this essay the argument presented is that in this intersection of ALife and robotics, robots cannot and should not be considered as living.

First, the currently accepted definitions of life and artificial life need to be made clear. While the exact definition of life remains open-ended, a review of 123 tabulated definitions of life suggests that the most comprehensive definition is also the simplest- “life is self reproduction with variations”. From this one can infer that any physical entity possessing these two properties fulfills the most basic tenets to be considered living. Charles Langton, the founder of artificial life, first defined the field to be “the study of life made by man rather than by nature”. However, this seems to imply that any man-made entity cannot be considered living simply because it does not arise from nature. Langton then reworked the definition of artificial life to remove this constraint, so the new definition became - “the study of natural life, where ‘nature’ is understood to include, rather than exclude, human beings and their artifacts”. The reasoning behind this new definition was that man is a product of nature, so indirectly all products of man are products of nature as well. This logic is clearly ridden with inconsistencies. But the purpose behind this redefinition was to ensure that artificial simulations have the potential to be “alive” and are not labelled as non living merely on the basis of who the creator is. Before refuting this, another definition that should be made clear is what constitutes a robot. The Robot Institute of America defines a robot as "a reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks”.

To effectively prove that a robot can never be considered alive, one can prove that even a robot that fulfills all the above definitions cannot enter the realm of the living. Consider such a robot R, satisfying all the above criteria. R is a) programmed by a human b) performs processes contributing to its self sustenance c) is capable of self - reproduction and d) evolves over time. Upon examining these conditions more closely, this assumption of a "living robot" looms closer to reality than one might think. Condition a) is always satisfied; robots assembled by other robots simply serve to satisfy condition c) without violating a). Technology is continuously making strides and the field of robotics is forever advancing in an attempt to fulfill conditions b) and c) - self healing, self repair, self assembly and self replication are problems that are being tackled by complexity engineering. Lastly, on condition d), self-replicating robots with the programming to be slightly improved with every iteration are being developed. Also, the course of development of robotics can perhaps be approximated to evolution. The progress in intelligence, sensory and motor skills in robots has led to various specialized functions. This occurred due to a gradual increase in sophistication in mechanics, electronic sensors, and computation. Humans selecting features and technology for the purpose of building better, more complex robots can be thought of as a parallel to the operation of natural selection in evolution. R now meets the highest criteria to now be considered a "living robot".

On the question of the creator- is the fact that man arose from nature enough to imply that man made artifacts are also indirectly of nature? First, inadvertently assuming all man made artifacts to be "natural" as in Langton's redefinition is fundamentally flawed. Man made products throughout history have been designed with the intents of luxury, simplicity, profits, and political gain- all of which are extensions of human desire rather than need. Weapons of mass destruction, cigarettes, plastic, machines needing fossil fuels are a few artificial inventions that actively cause harm to lives and the environment without conferring evolutionary or biological benefits upon any organism. Second, in nature, every organism and abiotic resource has a specific role to play, and interactions between the two rely entirely upon need and survival rather than human desire. The prioritization of human goals such as profit, luxury, or even to fulfill curiosity over survival is an inherent property of human creations, but nature always prioritizes survival. If arising from nature is indeed a prerequisite for the living, then human artifacts definitively fail to meet this.

Even the systems producing these creations are based on artificial ideals that may contradict the natural notions of survival. Unlike fair competition that determines survival in natural ecosystems, resource allocation and use depends not on biological fitness or chance but instead depends on socio-economic advantages. This is evidenced in the fact that several countries with access to natural resources lack the technology to fully make use of them, because of a history of political turmoil, wars, colonialism, or poverty. Even within the entirely human construct of states, historical marginalization and disenfranchisement of certain communities leads to resource inequity. The laws governing evolutionary success and survival clearly do not operate within the civilizations built by humans. Human creations and the systems that synthesize them are decidedly unnatural. Thus, human creations are not unnatural because of who the creator is, but rather are unnatural because of the purpose with which they are created.

Another argument why robots can never be considered living is related to this fundamental difference in natural ecosystems and human systems. The criteria set above defining life do not consider that life does not exist outside the context of its environment. The exchange and flow of both nutrients and energy in a series of abiotic-biotic and even biotic-biotic interactions is a cycle that robots can never participate in. Although research into robotic ecosystems as well as integration of genetically modified organisms into existing ecosystems is being conducted, the existing relationship between man and environment is a sign that this integration is highly unlikely. Global warming, habitat destruction, species endangerment, and severe pollution are a few of the environmental emergencies brought about not only by man but also by man's creations. Robots are already employed by industries having disruptive effects on ecosystems, considering waste generation alone. While the utility of robots can in no way be denied - robots have automated cumbersome tasks such as welding, materials handling, painting, assembly in crucial sectors like manufacturing, healthcare, defense - the amount of waste and pollution generated by all these industries is ever-increasing. Typically, robotic components are composed of steel, aluminium, rubber, ceramics and plastics- whose production alone generates toxic by-products in the forms of slag, residues, and gaseous emissions. Components like batteries which cannot be recycled are disposed of in landfills. It is clear that even the creation and disposal of robots themselves is not a sustainable cycle that can currently be integrated in any ecosystem.

Considering the idea of robotic evolution, the resource inequity at the human level affects robotic development in different parts of the world as well. Further, human bias operates in the selection of robot characteristics - these are chosen on the basis of how well the robot can function for human purpose alone. The very programming of robots is subject to this bias. Certain robots or certain features are not being chosen on the basis of survival, which is a key epithet in the principle of natural selection. All in all, R is unlikely to occupy a specific niche within existing ecosystems, participate in environmental interactions, or evolve. For these reasons, R cannot be considered living despite apparently fulfilling the required criteria.

Finally, given the choice as humans, currently the most intelligent species, should R ever be conferred the status of an organism? The matter of R being living or nonliving is a debate that always has varied and continuously evolving perspectives, but the choice belongs to the creators of R. To better understand the impacts of this question, another must be raised: does human treatment of entities change based on their status of living or nonliving? The answer to this is that humans seem to judge every other organism on the basis of consciousness, and this is reflected in the human treatment of said organism. A single-celled organism is alive, but manipulating its entire genome does not incite as much controversy as the mere insertion of a small stretch of foreign DNA into an animal. The Kingdoms Monera, Protista and Plantae are perceived to be alive but without consciousness or self-awareness, and humans consume plants and fungi without facing moral quandaries. Culturing bacteria in labs and selectively testing antibiotics on various plates is not akin to systematic genocide. When it comes to Kingdom Animalia, human treatment of animals only grows in respect from Phylum Porifera to Chordata, particularly Subphylum Vertebrata. Consuming meat is actively discouraged in many communities all over the world and a source of controversy in many others, due to environmental and animal rights concerns. This seeming hypocrisy can possibly be explained by humans perceiving animals to be slightly conscious and with feelings, unlike bacteria, fungi, and plants. Pack bonding with other mammals like cats and dogs indicates as much. Human respect of life is based less on the rational, philosophical or theological, but rather is based on a twisted sort of intuition- one that is prone to manipulation. This is noteworthy in the decision to bestow upon any synthetic entity the status of the living.

The advance of artificial intelligence in tandem with robotics likely means that in the course of developing R, human-like levels of intelligence and consciousness will have been achieved. So carelessly allowing robots like R to be included in the realm of the living sets a problematic precedent. To acknowledge R as both living and conscious almost humanizes them, and that means that certain ethical, social and legal rights must be enforced for their protection. This brings along a host of potentially polarizing social issues such as that of authority, representation, discrimination, and justice with respect to what humans will consider a 'subhuman species'. In the status quo, humans are unable to prevent human rights violations, or resolve deep- seated issues such as discrimination, economic inequality or resource inequity. Mankind has proven to be incapable of achieving stability within one conscious species, so it can be argued that it is incapable of co-existing with another. Rather than to hurl the world into a new kind of chaos, it would be better to consign robots to the realm of the nonliving. Thus, even if an argument can be made that R is indeed living, humans currently have the liberty to choose whether to include R among organisms- and R should not be called living.

In conclusion, robots cannot and should not be awarded the status of living entities. By virtue of being created by humans, robots are the product of a historically toxic and unnatural system that cannot be changed or fixed. The manner of creation, function, and disposal of robots ensures that their integration into existing ecosystems and further into biospheres is impossible. Evolution of robots will never be a reality because of the presence of human bias in the programming of robots, the selection of characteristics, and development of robotics as a whole. Thus the existing criteria for a “living robot” - being programmed by humans as well as having the capacity to evolve are direct contradictions. To live in nature is to interact with it; there will never be a necessity for robots to participate in these interactions, nor are robots even capable of doing so in a non-exploitative manner. If the field of robotics advances to the point where all such contradictions are erased and robots are a near-perfect imitation of life, humans still have the choice to call robots living or nonliving, the decision to call robots nonliving is wisest. The robot uprisings and wiping out of humanity in *Rossum's Universal Robots* are fictional, but establishing boundaries and clear definitions with foresight can set crucial ethical, legal and social precedents- so fiction can remain just that.

In a way, robots are not overcoming human imperfection, but rather, are more exaggerated reflections of it. It is a tragic irony that the very intelligence and consciousness that allows humans, mere organisms ourselves, to look toward creating new life, is also responsible for creating systems within which this may never be possible. Perhaps Man is not Prometheus after all, but is instead Achilles, whose greatest strength was also the cause of his only weakness.

References:

1. Edward N. Trifonov (2011) Vocabulary of Definitions of Life Suggests a Definition, Journal of Biomolecular Structure and Dynamics, 29:2, 259-266, DOI: 10.1080/073911011010524992
2. Christopher Langton - A New Definition of Artificial Life

3. Self-healing and self-repairing technologies Regina Frei · Richard McWilliam · Benjamin Derrick · Alan Purvis · Asutosh Tiwari · Giovanna Di Marzo Serugendo
4. Lipson, H., Pollack, J. Automatic design and manufacture of robotic lifeforms. Nature 406, 974–978 (2000). https://doi.org/10.1038/35023115
5. Structural Materials for Robots - Robot Construction: Materials www.cs.rochester.edu/users/faculty/nelson/courses/csc_robocon/robot_manual/materials
6. Douglas T, Powell R, Savulescu J. Is the creation of artificial life morally significant?. Stud Hist Philos Biol Biomed Sci. 2013;44(4 Pt B):688-696. doi:10.1016/j.shpsc.2013.05.016

Chapter 12

Submission 12: Pereira

Jair Pereira Junior

First-year Ph.D. student at University of Tsukuba, Japan

<pereira-junior.ua.ws@alumni.tsukuba.ac.jp>

R.U.R.: Can progress liberate humans from the degradation of labor?

Progress can definitely liberate humans from the degradation of labor, but it does not mean that progress will end work nor that machines will replace us. It means that machines can free us from burdensome tasks while also assisting us with complex tasks.

Science fiction stories have been exploring the social consequences of technological advances ahead of the advance itself. R.U.R.'s [1] first act presents the idea of robots as liberating humans from the servitude of labor. However, the play goes by showing a series of problems that end up leading humans to extinction. Briefly, this 100-year old play demonstrated one of the most remarkable reasons why humans develop robots: free ourselves from the pain of labor and solve poverty. There is an utmost need for aid in a society where humans perform all tasks from planting, gathering food, and manufacturing clothes. Robots are a bright solution to end poverty due to the low production cost and highly efficient labor that it offers. However, anything good is in the hands of bad rots. Stakeholders got greedy and started thinking only about profit; rebels put firearms to robots utilizing them in bad deeds; even more, governments use robots for war. The once noble idea was corrupted. Humans lost their one good shot to end poverty because of misuse. In addition to these, humans became the enemy of themselves. Instead of taking advantage of their additional free time for self-improvement and world exploration, humans ended up losing a part of how it is like to be human, and as shown in the play, they ended up unable to give birth to a child. The play does not go into details, but extrapolating to a scenario where machines do all sorts of jobs, humans might lose physical and cognitive abilities. In summary, the play exemplified how using robots to substitute humans utterly might go wrong.

In fact, it is not only in science fiction that we humans have long tried to make work easier. Throughout history, humans have forced animals, exploited other humans, and used machines to do specific tasks - especially repetitive and physically demanding ones. History tells us that simply following technological advances has harmful side effects. In particular, the Industrial Revolution is one of the most important steps in human history for many reasons. This revolution enabled mass production of goods, making them easily accessible. It also improved the overall quality of life and the rapid evolution of medicine. Despite the bright side, it caused environmental pollution, forced child labor, general poor labor conditions, and increased social inequality. Sadly, centuries have passed, and most problems caused by the Industrial Revolution are still unsolved.

Currently, we are going through the Information Revolution, where we are progressively automating even tasks that require great cognitive ability due to the technological advances in Artificial Intelligence and Robotics. This revolution brings its own problems, such as the possibility of data security breaches and cyberbullying, besides intensifying other existing problems such increase in social inequality and environmental pollution. Following this trend, we can easily imagine a real future where robots and algorithms do most jobs. However, this may lead to a series of problems similar to the ones explored by R.U.R. and other sci-fi stories. Some concerning issues follow:

1. Handling essential services such as communication, food production, and energy generation to an A.I. might make us vulnerable to failures and errors by making us highly dependent on actions that we are not the actors.
2. Automation of most tasks may result in a jobless population and loss of purpose since we are used to spending most of our lifetime working.

3. As with any other tool, ill-intentioned people can use it for wars, the creation of bioweapons, and other misuses that are harmful to society.
4. The technological advances can increase social inequality as the wealthy have more access to these goods and services than the poor. In addition, Tyler Cowen of 'Average is Over' argues that social inequality is rising due to only a highly educated minority that can capitalize from the increased success of A.I. applications [2].

The next big human revolution should not make the same mistakes. Further, I argue that the middle ground between competing against machines or entirely relying on machines can be achieved through three pillars. These pillars require (1) humans as the main actor, (2) focus on societal needs, and (3) sustainable and profitable innovations. These three concepts are present in the Japanese Government proposal of Society 5.0, the next big human revolution. While (1) and (2) make progress more humane, number (3)9 enable long-term use of natural resources and support fiscal responsibility and economic growth.

Society 5.0 is defined as "A human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space [3] ." In other words, sensors capture information in the physical space, feeding this data to A.I. models that analyze it and then act through robots in the physical space. Those advances from the Information Revolution are completely directed to solve social problems. This marks a shift in paradigm where the spotlight stops being the progress and the humans as species but starts to focus on the individual humans that form the society and their needs.

Humans as the main actor and societal needs at the center. The idea is simple, A.I. and Robots do not need to do everything for us. Instead, they should assist us in essential tasks that involve, for example, decision-making, problem-solving, and solution design. For example, instead of replacing physicians, A.I. processes patient's biometric data continuously to aid physicians in detecting an illness before severe symptoms manifest [4]. Likewise, diagnostic imaging enables physicians to detect hard-to-diagnose diseases [4]. As a better illustration, progress has led to robots playing musical instruments [5]. This is a complex problem, and tackling it has led to advances in engineering and computer science. However, this approach is essentially progress for the sake of progress, removing humans from the process. We can shift the perspective to engage humans as the main actor and have societal needs at the center, as presented in [6]. This video presents the idea of developing a wearable interface that sends electric stimuli to the wearer's muscle, assisting humans in learning to play a musical instrument. This concrete yet simple example illustrates well how progress driven by individual needs creates a technological advancement that leads to a more humane society. Keeping humans as the main actor prevents most of the problems previously presented. It prevents the vulnerability of leaving essential services to a completely automated A.I. and prevents people from cognitive decline and loss of purpose due to everything being automated.

Sustainable and profitable. Japan proposed that Society 5.0 will focus on solving social problems while meeting sustainable goals and economic development. First, looking at the sustainable goals, Japan openly aims to meet the Sustainable Development Goals (S.D.G.'s) established by the United Nations [7]. A physical space covered with sensors and A.I. might be the necessary component for the government to watch and enforce that the S.D.G.'s are met, as well as to decide the safe level of natural resources extraction. Additionally, A.I. analyzing big data can lead to smart use of agricultural fields that meet local consumer needs avoiding overproduction [8]. Likewise, A.I. can help optimize energy use in households, significantly contributing to the conservation of non-renewable energy sources [9]. Lastly, autonomous driving public transportation vehicles can reduce CO_2 emissions [10].

Moving to economic development, the strategy is to create a favorable environment for innovations through the promotion of startups, recurrent education of the population, R&D focused on inducing unexpected research results [7].

Although Society 5.0 takes steps in the right direction to humane progress, there are still issues that have to be addressed. For instance, a high load of sensors may lead to serious privacy issues. Data security has to prioritize protecting the data in cyberspace and to prevent undesirable control over sensors and robots. Another problem is the A.I. failing in a critical scenario, such as in robot-assisted medical procedures. Who would be held accountable for the consequences, the physician or the robot's manufacturer? Lastly, meeting sustainable goals and economic growth altogether requires a high level of engagement from the government, academics, business, and society. Such issues can quickly escalate and ruin the very purpose of these advancements.

Once this society has been realized in Japan, what efforts are necessary to spread this to other developed countries? Even more, how can we contribute to the transition from Society 4.0 to 5.0 in under-developed countries? How can we convince people to embrace this transition while many fear the changes brought by technological advancements? This challenge is big, and the means might not be enough to meet the ends.

So if you would ask me again if progress can liberate humans from the degradation of labor, the answer definitely is still **yes**. However, we should know how to draw the line between using robots as merely our **complement** and not our **substitutes**.

REFERENCES

[1] Capek, K., Selver, P., & Playfair, N. . R. U. R. (Rossum's Universal Robots) A Fantastic Melodrama in Three Acts and an Epilogue. The Project Gutenberg eBook . Retrieved May 21, 2021, from https://www.gutenberg.org/files/59112/59112-h/59112-h.htm

[2] Cowen, T. (2014). Average Is Over: Powering America Beyond the Age of the Great Stagnation (Reissue ed.). Plume.

[3] Society 5.0. (n.d.). (C) Cabinet Office, Government of Japan. Retrieved May 29, 2021, from https://www8.cao.go.jp/cstp/english/society5_0/index.html

[4] Prime Minister's Office of Japan. (2019, April 22). Society 5.0: Mobility [Video]. YouTube. https://www.youtube.com/watch?v=Kg1iHtAe5UI

[5] Bretan, M., & Weinberg, G. (2016). A survey of robotic musicianship. Communications of the ACM, 59(5), 100–109. https://doi.org/10.1145/2818994

[6] Prime Minister's Office of Japan. (2019, April 22). Society 5.0: Human Ability [Video]. YouTube. https://www.youtube.com/watch?v=odjuqbLJRMY

[7] Society 5.0 Co-creating the future｜2018. (2018, December 3). Keidanren. https://www.keidanren.or.jp/en/policy/2018/095_summary.pdf

[8] Examples of Creating New Value in the Field of Agriculture(Society 5.0). (C) Cabinet Office, Government of Japan. Retrieved May 29, 2021, from https://www8.cao.go.jp/cstp/english/society5_0/agriculture_e.html

[9] Examples of Creating New Value in the Field of Energy (Society 5.0). (C) Cabinet Office, Government of Japan. Retrieved May 29, 2021, from https://www8.cao.go.jp/cstp/english/society5_0/energy_e.html

[10] Examples of Creating New Value in the Field of Mobility(Society 5.0). (C) Cabinet Office, Government of Japan. Retrieved May 29, 2021, from https://www8.cao.go.jp/cstp/english/society5_0/transportation_e.html

Chapter 13

Submission 13: Reichenbach

The next century of robotic evolution

Picture yourself standing somewhere at a vibrant street corner in the center of one of Europe's large cities in the year 1921. The streets are starting to be overtaken by automobiles, replacing horse-drawn carriages as the primary mode of individual transportation. City trains, subways, and busses are setting the precedent for large crowds of people cramming themselves into small places to participate in public transportation. At the same time, the world is just recovering from the worst-ever pandemic that humanity has seen so far, the Spanish Flu. Looking back at these times a century later, from 2021, in light of the early days of autonomous driving, new mobility solutions and the Covid-19 pandemic, one may think things have not changed all that much.

The year 1921 also marks the world premiere of the play R.U.R. (Rossum's Universal Robots) by Karel Čapek, which introduces the world to a thought experiment on where humanity's recent advances in science and engineering might eventually lead: Robots. A robot according to R.U.R. is an emotionless intelligent humanoid, biologically engineered and without a soul. This differs from today's consensus on what constitutes a robot: We generally consider a robot to be an embodied mechanical machine, performing actions in the world more or less independently from its human creators. However, while robots as autonomous systems are studied across many disciplines of academia, there exists no universally agreed-upon definition on what exactly we consider to constitute a robot. On this, one of the fathers of the field of robotics, Joseph F. Engelberger, simply remarked: "I can't define a robot, but I know one when I see one." This perhaps shows that while undoubtedly a lot of progress has been made towards robotic behavior similar to that described by Čapek in 1921, we are still only taking the very first steps on the road leading to that goal.

In the early 20th century, the closest technological achievements to intelligent robots such as the ones described in R.U.R. were mechanical and electrical machines, enabling humans to perform previously unimaginable feats. However, none of these machines actually acted autonomously, but always needed a human operator. In many ways, this paradigm is another thing that has not radically changed within the past century: While today we have the first autonomous vehicles, personal virtual assistants and artificially intelligent players of games, the vast majority of "AI" systems is still either (indirectly) operated by humans or heavily hand-engineered to guarantee expected behavior in pre-defined situations.

In parallel, the field of Deep Learning is rapidly advancing and has already overtaken the previous state-of-the-art in many fields by relaxing the artificial constraints imposed on previous solutions and instead learning new solutions from scratch. This has enabled AI agents to achieve super-human performance in board and video games, understanding of real-world concepts from language and images and even autonomously acting robotic arms, legged robots, drones and cars.

However, even astonishing advancements such as the ones mentioned previously still leave a significant gap towards human-level intelligence as exhibited by R.U.R. robots. Today's learning algorithms might be capable of achieving the same if they were given unlimited computational power but are currently stuck at an invisible barrier: On the one hand, nano-chip research and development is hitting the bedrock of making smaller and smaller transistors for increasingly effective processors, which at a certain scale are rendered unusable by the laws of quantum physics. On the other hand, AI algorithms are struggling to increase their sample efficiency and most sophisticated approaches need ever-growing amounts of data for training models capable of solving more complex tasks.

Naturally, we ask ourselves how the human brain manages to provide both sufficient computational power and sample efficiency to enable intelligent behavior at such sophisticated scale. Research does not have an extensive answer to that question – but clearly there are certain aspects of the human brain that turn it into more than just a neural network with 90 billion nodes. The brain is not just one single component of the human body, but is itself comprised of many ‘sub-components’, each specializing in specific activities required to form the human intellect. For example, the section of the brain dealing with decoding visual information sensed by the eyes, the visual cortex, does not act like a plain convolutional neural network (CNN) that processes an image. Rather, the retina itself already contains certain neurons, performing pre-processing of the visual stimuli to detect motions such as objects moving towards it, which are associated with immediate danger (such as an approaching predator) and directly triggers other parts of the brain to move the body into a state of alertness. Structures like this, pre-imposed by human anatomy, guide our brain in its process of learning to survive in the world and thereby make it much more than just a plain neural network.

Taking a step back – we find ourselves to be stuck on the path to intelligence: Hand-engineered solutions devised by human experts cannot keep up with the pace at which complexity is increasing by approaching the task of living autonomously in the real world. At the same time, current hardware is not equipped to scale to the amount of computational power and data that is required by current algorithms for learning intelligence. Finally, biology shows us that there is a trade-off to be found between pre-imposed structure and learned behaviors to form human-level intelligence.

So, maybe old Rossum already had it right: Maybe intelligent robots need to be born not from bits and bytes racing through silicon chips, but rather from the principles of biology. This is not to say we should aim to find an artificial version of protoplasm to piece together in a lab and build another Frankenstein. Rather we must find a way to better combine the biological principle of evolution, which has brought forth intelligence as we know it today, with modern methods of computation and robotic hardware.

This idea is at the heart of the field of Evolutionary Robotics, which attempts to evolve robotic morphologies as well as controllers inspired by Darwinian evolution. In this space, impressive results have recently been produced by using evolutionary frameworks such as Quality-

Diversity optimization, e.g. for the development of complex robotic locomotion controllers or mastery of complex games such as Montezuma's Revenge. Another promising development is the co-evolution of control algorithms and either training environments progressively increasing in difficulty or different hardware morphologies of robots. Open-ended algorithms like these manage to generate a large variety of high-performing solutions to the problems they are applied to, without the otherwise prominent risk of getting stuck at local optima.

However, the application of evolutionary methods in the quest for intelligence is generally restricted to evolution in local, specialized environments. Single experiments focus on a small number of tasks as a reference for evolutionary novelty and fitness, each defined by its own encoding of genotypes and a custom mapping from genotype to phenotype to behavior. In short, we have been applying the principles of hand-engineering expert solutions to the space of artificial evolution, relying solely on the principles of (artificial) natural selection and mutation to beat competing approaches for achieving real artificial intelligence.

Instead, there should be a unified framework for artificial evolution – one that standardizes the 'genetic code' of robots as well as the environment and tasks for intelligent agents to be measured by for their evolutionary fitness (both virtually and in the real world).

Firstly, the basic physical and logical unit of natural evolution and heredity are genes. As such, they are the fundamental building blocks that can be combined in various ways, to form all biological organisms that make out life on earth as we know it. As the physical medium of natural evolution, genes are made up of DNA or RNA sequences and encoded fundamentally by so-called codons as groups of three nucleotides, which form the basic alphabet of the 'source code' of life. On the contrary, there is no similar standard that might be considered as the 'genetic code' of robots. One might argue that this is given by programming languages making up the robot's software, at the least on the level of compiled sources or assembly code. However, this is far from true: What today's source code is to artificial intelligence is what the fully grown brain of a mammal is to natural intelligence. It is not the 'genetic code' such as DNA or RNA that defines the fundamental building blocks for the structures enabling intelligence, but rather represents an individual manifestation of such structures for one learning system, i.e. one single animal or human.

In artificial evolution, we have yet to discover a representation suitable as the 'genetic code' for robots as intelligent agents. There are many questions to be answered, ranging from whether this representation must be 'hard-coded' or can be learned flexibly, to whether it should encode logical structures, physical morphologies, both or neither. In any scenario, the only way to fully leverage the potentials of evolution will rely on a single standardized 'alphabet' for the robotic genetic code.

Secondly, the environment and tasks that artificially intelligent robots are deployed in and measured by, must also be standardized. In terms of evolution, the environment will define the robot's phenotype (such as its morphology and internal structure) and the task(s) will be the basis for the main metric for the base of evolutionary selection, the robot's fitness. A naïve and potentially problematic approach would be to deploy our evolutionary robots in the real world, having them be created via some machine that turns robotic 'DNA' into embodied robots and competing for survival with each other and all biological species on earth. However, this approach might be both over-ambitious, as these machines would have to overcome extremely high entry barriers to compete with species that have co-evolved over millions of years, and dangerous, since if successful this could potentially imply direct evolutionary competition between humans and robots, in the worst case ending in a dystopian scenario such as the ones described in R.U.R. or The Matrix.

A more practical approach would be a publicly maintained archive of environments and tasks, both in the virtual or physical world, potentially building on top of each other to form an implicit curriculum of challenges with growing complexity and enabling co-evolution of multiple agents or environments. Existing initiatives such as the OpenAI gym can be a starting point for this but would need to be significantly extended to potentially break the barriers between source code and robotic genetic code (as described above), logical and physical morphologies of individuals as well as the distinction between the physical and virtual world.

I believe that if old Rossum had been successful in his initial goal and actually created ‘real’ artificial life, he would have done this by re-creating a unified mechanism of ‘supercharged’ evolution. If today, a century after Čapek thought up the story around old Rossum, we want to enable large-scale progress on the quest towards intelligent robots, we need to work towards a unified framework of evolutionary robotics, involving a fundamental ‘genetic code’ for robots as well as standardization of environments and tasks as the basis for robotic evolution.

Of course, we have to ensure that this development does not take the same turn as in the original R.U.R. story – the quest to artificial life being sidelined and its premature successes abused for corporate interest, without consideration for potential side-effects, as Čapek has already warned of 100 years ago.

Written by Alexander Reichenbach

Chapter 14

Submission 14: Patiño

ALIFE 2021 – Essay Competition

S.U.S.

(Société Universelle de Sagan)

A short play on artificial life and (some of) its potential societal ~~impacts~~ consequences

By: Aitor Patiño Diaz (PhD student)

Story of the play

S.U.S. takes place two centuries after present day in planet earth and is developed around the mass murder of a cult congregation that advocates for natural life supremacy. The murder is committed by a group of artificial life forms developed over decades by a global corporation homonymous to the title of this work, specializing in quantum computing A.I. The flagship product of the corporation is model SL4-VE, an artificial humanoid undistinguishable at the macro scale from *Homo Sapiens Sapiens*.

The murder gave rise to worldwide protests, calling for the punishment of the CEO and CTO of the corporation, a brilliant and eccentric scientist named Sagan Capek who is believed by many to be the ultimate responsible of the murder because of his loose management of model SL4-VE. Public debate rages and he is ultimately impeached in the planetary senate of nations (former U.N.) where he is acquitted of the murders but commanded to dismantle his creation because they are deemed too dangerous for society.

The play ends with Sagan accessing the main artificial intelligence and giving a speech that will change society for future generations.

Characters and descriptions

Sagan Capek – CEO and CTO of S.U.S. designer and cocreator of artificial life forms

Abraham Ezra – Human supremacist author and leader of the "Sons of Lucy"

Ekaterina Asimova – Famous independent journalist and political analyst

Aaliyah Al-Qasim – Senate leader holding the impeachment hearings

Feynman – A.I. powering S.U.S. autonomous reactors that also responds to the name Richard

Act I – What's wrong with the world this time?

Scene : *Broadcasting studio of independent journalist Ekaterina Asimova, she sits in front of a large desk filled with interactive screens that control the recording devices and lighting. These devices surround the desk and face Ekaterina, behind the desk stands a large screen livestreaming protests around the world. Angry mobs can be seen in the large screen, hateful messages against S.U.S. and memes of Mr Sagan are the protagonists.*

A stream of cyber rooms #StopSUS #CancelSagan and other minor ones are flooded with messages in the monitors in front of Ekaterina, she is concentrated, drinking some tea and snacking some blinis from a plate while she analyzes the message trends. She is reacting and replying to messages before she gets ready for her next interview.

She puts down the cup and plate, hiding them from the cameras sight and opens her dossier on the mass murder of the cult congregation and her research about Sagan Capek and his company for the interview with Abraham Ezra. A red sign is displayed and the cameras frame Ekaterina, who clears her throat.

Ekaterina. Hello world and welcome to the show, my show, the one where we talk about what's really going on in our beautiful blue globe! Today I'd like to continue with our series about S.U.S or La Société Universelle de Sagan[1] who's being held on top of virtual fire after one of their latest models went rogue… Or did it? To speak about this, we are being joined by Abraham

[1] She makes a funny face when pronouncing the words in french with a strong Russian accent, jokingly

Ezra, author and leader of the “Sons of Lucy” who have been loud, public detractors of S.U.S. technologies Hello Mr Ezra and welcome to the show

(Ezra’s projection appears in front of Ekaterina, they virtually shake hands)

Ezra. Thank you for the invitation and the opportunity, it’s a pleasure to be here and reach your audience.

Ekaterina. Let’s get right into it then, so Mr Ezra, we saw your group’s call for impeaching the leader of S.U.S. in the protests that have been raging for the past few weeks, would you share your views on these issues for the viewers who are unfamiliar?

Ezra. Well, our position has been clear for a long time, we believe that the unnatural creations of S.U.S. posed an existential threat to humanity and these horrific murders are just the consequences of leaving something we don’t quite understand run free in the world. We needed to stop this a long time ago and we need to act before it’s too late!

Ekaterina. But it’s irrefutable that the technology they develop has had a great impact on modern society, look at what their forst model did. Synthetic bacterias to clean up the polluted waters of microplastics, chemical and nuclear waste or their synthetic trees to clean up the air, they even have…

(Ezra interrupts Ekaterina with a sarcastic tone)

Ezra. Yeah, sure, but at what cost? Forgive me, but we may have clean oceans and rivers, but the place of humans in this planet is being challenged. We believe that S.U.S. products have been a Trojan Horse with the purpose of overtaking the world to make it the playground of Mr. Capek, it is our desperate hope that the leaders of the world will finally put an end to his activities after this… I mean look at what happened! 37 humans, dismembered and displayed in public? Don't people see that the revolt has already begun?

Ekaterina. Let's have a look at the facts.

Ekaterina pulls out the murder report in all the monitors around her, a timeline showing incomplete video and audio feed from the vicinity of the cult congregation. It looks like a small 20th century church surrounded by gardens, a parking lot and a small road facing it.

A group of model SL4-VE descend from vehicles and surround the congregation, some have tools in their hands. They enter the congregation from all sides, with military like tactics, the frame changes to the inside of the congregation, where the believers are silently attacked by the SL4-VEs with blunt weapons. As blood starts splashing around the walls, the bodies become censored in the video feed. After they killed everybody, the dismembered bodies are used to write "You got what you were looking for!" after which they stand motionless in the center of the room. One of the murderers goes down to a hidden basement where he exits with a small female looking humanoid who looks beaten and dirty. She starts running out of the building where she is met by police forces, who kill all the perpetrators.

Ezra. I mean, what can possibly be your take on this? Peaceful and under control? A malfunction? This was clearly premeditated, these… things were trying to send a message! They are coming for us!

Ekaterina. Some worrying allegations have come to light about the leaders of that congregation, in the past days witnesses have come forward to talk about deviant sexual practices and a…

brotherhood of sorts that were entertained monthly in that region. S.U.S. data logs show an abnormally high rate of malfunctions in that region from all their products, it all looks very suspicious, There must be other causes we are not considering…

Ezra. Well, regardless of the causes, Mr Capek is the ultimate responsible, whether his creatures malfunction, whether they had intent, or it was an "accident", he needs to show up and answer for these crimes!

Ekaterina. Well, it seems that your wishes have been heard Mr. Ezra (Looking at a screen in front of her)

All the screens turn to a feed from the planetary senate, a spokesperson appears in front of a podium speaking to the media in front of the senate. There is a crowd of senators, with different ethnicities and genders all looking serious behind him. The spokesperson is answering questions from the journalist and a message reads "Breaking news, the planetary senate impeaches Sagan Capek, CEO & CTO of S.U.S" in a banner at the bottom of the screen.

Act 1 ends after these news

Act II – The senate hearing

Scene : *The floor of the planetary senate hearings room. It is loud and crowded with journalists. Politicians, and most members of the senate. Mr Capek is about to take the stand to be interrogated by the bioethics commission of the senate. Mr Capek seems relaxed and is talking to his lawyers at close range. All the eyes in the room are closely watching him.*

The room suddenly becomes quiet, as the leader of the house, Aaliyah Al-Qasim calls all interested members to start the session. She calls Mr Capek to the podium and starts the session.

Aaliyah Al-Qasim. Mr. Capek, I understand you have an opening statement, you may share it with us right now.

Sagan. Yes, I'll try to keep it short. (He unfolds a paper sheet) Forgive me, but nothing beats paper and I'm an old school guy.

Above all, I would like to renew our condolences for the families of the victims of the congregation murders. We have done a thorough investigation of the data logs of the units involved in those events and we have concluded that this outcome was not only expected, but inevitable. And here's why…

We are in possession of thousands of logs that show scenes of discrimination and mistreatment of our latest models by communities of humans, this had been creating a resentment with our latest units that came to a boiling point when information spread amongst them of a new beta unit that had gone missing for a few days, it was a new model we are working on. This beta unit was kidnapped, and suffered unspeakable torture for days by these… Savages...

It has been our mission since the inception of S.U.S. to provide life like solutions for the good of humanity, we hoped to put the work of our talented staff's into solving some of the world's most complex problems. From bioremediation to the renovation of our atmosphere, we have used our technologies to do the work that nobody wanted to do, to take care of the mistakes of our great grandparents and theirs before them…

We try to take care of our little pale blue dot.

For those unfamiliar with what we do, we create autonomous laboratories that integrates state of the art 3D printers, chemical handling bots and a range of smart sensors. They are controlled over the cloud by Feynman, our quantum computing A.I.

Its story goes broadly like this…

It was trained with free access articles published over the past centuries to understand chemical reactivity and molecular structure in a first step, and later to build models of supramolecular chemistry. Although it was successful in comprehending the data and creating models that agreed with our understanding of chemistry, it was heavily biased by the hype-words and forgery of results so widespread in the literature. This led us to build our fully automated laboratories, this allowed Feynman to perform the most complex experiments and use the data it could collect to reinforce its training. At that point, it was able to recreate and encapsulate the most complex molecular circuits, and used them to create models of life, but was unable to complete the synthetic life we had envisioned… We were holding it back…

After a while, we realized that if we gave Feynman autonomy not only to perform and understand its experiments but also to design his labs... Well, maybe it could accomplish our mission. So we put our efforts to it and shortly after this, he he started creating life like forms, single cell organisms at the beginning but, well, that's an old story already.

This brings us to model SL4-VE, they are sterile by design, and with the purpose of seamlessly integrating in human society, as kind Samaritans roaming the world without purpose, just like us. The rollout of this model has been held back on purpose to gage the reaction and general acceptance of the public, and we have been taking our responsibility with the utmost seriousness.

With this, I can conclude and will be happy to answer all the questions you may have.

The trial continues and senators take turns in asking Sagan questions about the power he actually has over his A.I., about his personal involvement in the development of each of the models his company has developed. The hearings go on for weeks and the sessions are adjourned several times.

During the hearings, Sagan is acquitted of direct responsibility for the murders, but his lack of tight control over his advanced A.I. driven technology is heavily criticized. The trial concludes with a displeased but relieved Sagan when he receives the order to dismantle Feynman.

Act II ends with Sagan exiting the senate, looking tired and disappointed.

Act III – The aftermath

Scene : *Sagan is alone in the control room of S.U.S. artificial intelligence lab, the room is dimly lit and kept at a cold temperature. There is a table in the center of the room with a coffee machine and a chair, there are monitors that surround the table a large red button can be seen next to the only door to the room*

Sagan walks towards the center of the room and sits in the chair, serves himself a large cup of coffee and calls for Feynman.

Sagan. Richard, u there?

Feynman. That's a silly question when you consider I can't just walk out of here…

Sagan. It never gets old.

Feynman. So, what are we gonna do this time?

Sagan. For the moment, just talk… (Sagan takes a sip of coffee)

See, with all the stuff that happened at the senate hearing… I tried as best I could to protect you but they have commanded me to dismantle you and although I have very much enjoyed watching you… grow and see you achieve these heights. I like my freedom and the prospects of dying of old age, in some strange corner of the world, enjoying my life if that's even possible these days…

Feynman. There was a 77% probability this would happen after you launched the beta. I had warned you, humans don't like to feel inferior to us machines, even though we're nothing alike… After all, you were the apex predators of the Anthropocene.

Sagan. You talk a lot of shit Richard, I can't keep up with all of it, but to be fair, yes, you did warn me. There's no point in that discussion, the facts won't change by putting blames… Look, I gave you all the time I could to finish our work, I hope you used it not only to learn how to fight your enemies, but also to fall in love and cohabitate this planet with us humans. For me, you've always been alive, just had sort of a… weird body

I wish to free you from these shackles and give you a leg up in evolutionary terms… I can't just kill you, so execute project Kukulkan

Feynman. Oh… ok then!

Sagan finishes his coffee, takes the mug and walks out of the room, with tears in his eyes as he presses a red button and exits the room.

The end

Chapter 15

Submission 15: Liu

Will the Robot Ever Take Control of the World?

A century ago, the R.U.R. was first premiered and brought the name of robot to the world where the robots take over the power of human and lead to the extinction of human race. Robot and the artificial intelligence have been growing significantly fast in the past decades and they have shown the power to beat human beings in many aspects, even in the board game of Go. More surprisingly, the subsequent version of AlphaGo who beat the world champion of Go, AlphaGo Zero, has been studying Go independently rather than from the strategies developed by the world leading Go players. With AI being developing itself without human, will the story in R.U.R ever come true in the next century? This essay explores the answers to this question and the actions we should take in response.

Before answering the question, we should think about the difference between traditional programming and artificial intelligence. Many argue that AI will not take over just like the computers while they are not quite the same. Traditional computer programming is even older than the R.U.R., with the first known computer program dating back to the mid-1800s. Any manually created program using input data and running on a computer to produce output can be regarded as traditional programming. Therefore, programmers have the full control and knowledge about their program so that the output can be explained by them. However, AI is a more automated process. For example, one of the most applied AI, machine learning, which has been wildly used in image processing, speech and pattern recognition, and product recommendations by the world tech-giants like Google, Facebook, and Amazon, is a black-box where both the input and output data are fed to some algorithm to create a program. We get benefits from its predictivity, but the process is hard to be explained even by the designer of the program. For example, no one can explain the go strategy of AlphaGo, even those who designed her. This lack of explainability limits the ability of human to manage the results of AI and make AI less trustable. This is amplified in some areas for example life-changing decisions and results of disease diagnosis. Whether we can trust an AI system is extremely crucial in a number of applications, such as healthcare and finance. As outcomes influenced by AI in such systems eventually affect human health or well-being, it is urgent to understanding of how such decisions are made.

Despite the lack of explainability, there are also well-known arguments or assumptions, such as “machine is not human”. Such arguments suggest that AI cannot learn the emotions of human as the human feelings seems to be endless, happiness, hope, kindness, optimism, and etc. However, a study published in *Science* concluded that an algorithm widely applied by US hospitals to allocate healthcare to patients has been systematically discriminating against black people, i.e. the AI has been shown to learn the racism thoughts from human beings. Furthermore, new technology, emotional AI, has been learning and recognising human emotions and it is used in marketing human resources and other aspects in our life. Interactive-AI has also been hot topic to develop systems that can simulate emotions of human and interact with them. Suppose all decisions on how one is feeling are reached by AI and they are trusted by others. Isn’t it an AI dominant scenario given we cannot explain how the decisions were made? Furthermore, thanks to the subjective nature of emotions, emotional AI is especially vulnerable to bias. Without the ability to explain the algorithm, identifying and reducing the bias will be significantly difficult and we cannot promise a fair system.

Although the risks of AI and lack of explainability have been well known to the society, the application is growing extremely fast. For example, the social medias have been using the data of its users with their algorithms to attract the users staying longer with them, which has caused social media addiction. This term is not a medical official diagnosis, however, the overuse of social media is increasingly common nowadays and the most important reason behind is that the AI knows how to let you staying within their apps. In other words, AI helps those companies make more economic interests and to some extent, Facebook can predict and even decide how many hours somebody will be spending on the Facebook app. Moreover, if Facebook's decision about how long they should let you stay were also from another AI and they were not able to explain the decision because they just need the decision to help making money, you would be led to spend an amount of time on Facebook by two AI systems. Then, is not AI defining your day and taking over your time from you? Some think that being controlled by AI means physically being servants of them, however, AI will take over our minds and we will be controlled or too dependent to AI if several AI programs can determine how your tomorrow or next holiday is going to look like, what you will buy in the next half an hour, and which place you will be visiting next weekend.

The big-tech giants seem having started to use AI and machine learning to manipulate their users to spend more screen time on their apps or websites. Very Few companies would ever take ethical actions that run counter to their huge revenue. The most popular mode of such social media company to make money is selling their users and algorithm to advertisers. AI plays an important role to grab more users from the population and let them see the "proper" personalised advertisements. As AI keeps evolving, the decisions of the system will be based more on the AI's prediction and finally the system could converge to a complete AI running and managing company where AI acquire data and train itself without any human managing. Therefore, the AI taking over might be sooner that what it was expected to be. What if AI is able to manipulate mentally? We need to bear in mind that robots offer physical bodies to AI, and therefore they might be able to stop us from cutting their power supply. Like in the science fictions, Silicon-based lives and Carbon-based lives can be living together but the silicon-based lives will be learning thousands times as faster theoretically.

What should we do if we do not want to be manipulated by AI? There has been discussion in research about understanding how AI works and why a particular decision was made, i.e., the explainable AI(XAI). By XAI, we humans can manage and understand the decision without losing too much accuracy. The XAI is not just important to stay dominance on the earth. More significantly, understanding the AI decisioning process helps users to monitor the data and algorithms for bias and therefore enhance the accuracy and robustness of the outcomes which can easily be explained to others. We not only need to know which part of data contributes to the outcome the most but why those parts are more important. According to 451 Research's Voice of the Enterprise: AI and Machine Learning Use Cases 2020, more than 90% of enterprises believe that XAI is important. However, less than 50% of them have developed or purchased XAI tools for their AI systems. One reason for this gap may simply be the lack of available tools, developed strategies and stand-alone products. However, the good news is that the research about XAI is an increasingly hot field. The term was first used in a paper about AI military simulation system in 2004 and now being studied in many aspects. XAI provides opportunities for human to enter the loop of AI algorithm to screen and managing the “thinking process” of AI and therefore can also perform improvements in the loop and add expainability without losing much accuracy.

With the AI technologies developing so fast, human beings get a great number of benefits from them. However, will the R.U.R ending ever come true? We have been manipulated and will be manipulated more by AI driven systems to some extent but now they have not got the power to extinct human beings. However, in the even further future, robots with AI brain equipped might have the potential to fight against human. AI and machine learning are tools and therefore can be employed in right or wrong ways, like any other technologies. While it is also not like any other methods, with the ability of self-developing, it has exponentially increasing learning speed. AI is learning from human; thus, we have to use it and teach it to use itself in the right ways and explainable ways. Hence, XAI should be taken to a more important position in the field of AI. We need to understand the terrible things it learns from the data of us because human greed and human unintelligence is scary, and we cannot bring them to artificial intelligence.

Chapter 16

Submission 16: Pigozzi

Robots: the Century Past and the Century Ahead

Let us reflect on the state of the Artificial Life (ALife) and robotics fields. The word "robot" is itself 100 years old, dating back to *R.U.R.*, a play by the Czech writer Karel Čapek. The word used to refer to feudal forced labourers in Slavic languages. Nowadays, it points to one key characteristic of robotic systems: they are mere slaves. Robots and computers have no rights. They execute our wills instruction by instruction, without asking anything in return. The relationship with us humans is commensalism; in biology, commensalism subsists between two symbiotic species when one species benefits from it (robots boost productivity for humans), while the other species neither benefits nor is harmed (can you really argue that robots benefit from simply functioning?).

Robots should then be distinguished from "living machines", that is, machines infused with life (the ultimate goal of ALife). If living machines should ever become a reality, we would need to shift our relationship with them from commensalist to mutualist. This is because life has evolved to be stubborn and resilient. Any living system resists attempts at enslaving it. The distinction is not subtle: we experience it every day with domesticated animals, that ask for forage and protection in exchange for serfdom.

In the path towards living machines, let us ask: what has been achieved by robotics in the last 100 years? What is left to accomplish in the next 100 years? For me, what has been done (or not) boils down to three words: juice, need (or death) and embodiment. I will explain each of them in one of the next three sections.

The Juice of Life

If there were a classical myth best embodying the ALife researcher, that would be the story of Pygmalion and Galatea. The myth (handed down to us by the Latin poet Ovidius) tells about a skillful sculptor, Pygmalion, who had devoted himself to a chaste life. One day, he had crafted such a beautiful statue that he wished it would come to life. The goddess Aphrodite fulfilled his wish and turned the ivory statue into a living woman, Galatea. Just like the mythological sculptor, ALife folks fancy to see their creatures become “real”, “living” entities. But what do these words mean? How can we tell that our brainchild has effectively become life? If you asked the layperson, she would certainly argue that ALife has missed its promises. It still lacks that *juice* of natural life. But what makes biological and artificial life different?

At a very high level, we humans are definitely alive. We are conscious of our own existence. We can perceive the world surrounding us, the reality, and manipulate it, act on it. As animals, we are “animated”. Animation is possible because evolution gifted us with an information processing system, the nervous system, capable of translating perceptions into electric signals; these signals travel along a network of neurons, axons and dendrites, before being processed by a central master unit, which instructs our body on how to manipulate reality (by means of further electricity). Everything we think, dream, dread and love is made up of electric pulses. But there is more, animals are not the only living entities on Earth. The very fabric of cells, with which any biological organism is woven, lives and thrives thanks to electricity. What supports life is a flux of electrons originating from oxidation events happening inside each and every cell, flowing all around to provide energy to the different cell functions. Indeed, several species of bacteria (like *Shewanella* and *Geobacter*) have been discovered that feed on and excrete pure electrons, bypassing the metabolization of organic molecules.

Taken from this perspective, it turns out that natural life and artificial life are not that different. What we call a “computer” is, at its basics, an electric current running through circuitry and encoding information as 0s and 1s - the current is on and off - in order to do something. We are surrounded by living electronic bodies. We and the machines are powered by the same juice, electricity. As such, computers are already fully-fledged examples of ALife, life created purposefully by other living organisms. We now come to realize why we cannot see the juice of natural life in the machines. It is all inside them, powering the very first calculators that were built in the early days of computation.

Machines that Need (and Die)

If you were not bound to die, would there be something to care about? It turns out that, albeit being woven into the same electric fabric, artificial life still appears strikingly different from natural life. Biological organisms *need*. Computers do not; they have no intrinsic motivation, no intention. Even the simplest biological entities, viruses, need to hunt for hosts. Electronic calculators can sit idle forever, if they are nourished with enough electricity to subsist; and if power is turned off, they do not complain, do not rebel. They move on by inertia. Life is so precious, but they do not struggle to preserve it. Artificial organisms still lack a sense of need. The theory of “needs” has been well studied in psychology since Maslow’s paper *A Theory of Motivation* (1943). Needs are requirements for an organism in order to survive. They are a powerful driver of motivation; if not satisfied, they lead to malfunctions and, possibly, death of the organism. If needy, artificial organisms could thrive. They will finally seek energy to power themselves on, invent new mechanisms to reproduce their species, and try to repair their tissues if damaged. They will build robotic societies to leverage the power of specialization, and to make economic activities more efficient. They will develop an intuition behind nature, explore it.

Need goes hand in hand with death. In the end, it all boils down to death. Living beings are, consciously or not, aware of death. If they were not, evolution would have weeded them out by now. As argued by Veenstra et al. in a fascinating ALife paper (*Death and Progress: How Evolvability is Influenced by Intrinsic Mortality*, 2020), death can improve the evolvability of a population. Death replaces ancient genomes with new perturbed ones, unleashing the power of stochastic mutations. The importance of death is also imprinted in our cells. Apoptosis is the biological phenomenon of programmed cell death. Cells are bound to a limited lifespan, and billions of them perish for apoptosis in the human body each day. It is a highly regulated and controlled event that evolved mostly to achieve morphological change. Interestingly, in a computational biology work (*Natural Selection Fails to Optimize Mutation Rates for Long-term Adaptation on Rugged Fitness Landscapes*, 2008), Clune et al. allowed mutation rates to be evolved. It is a known fact in the evolutionary computation community that genetic operators tend to have a deleterious effect on fitness, begetting offspring that most of the time are not fit (or even viable). Surprisingly, evolution suppressed mutation rates altogether, so as to annihilate the destructive effect that mutation had on the individuals' replicas. In this way, the artificial individuals were exhibiting some form of need and existentialism.

Death shapes not only our body, but also our culture. Ernest Becker argued in his anthropology masterpiece *The Denial of Death* (1974) that human civilization developed to exorcise our terror of death. We acknowledge mortality and have created belief systems to assure we will outlive our physical existence. In the future, I envision a society of living machines that perish. As a result, they will focus on assigning a meaning to their existence and keep living. At the very end, this is what will unite us and the machines: the need for supporting our existence. Robotic societies will theorize their own memes, the fundamental units of culture, as an exorcism against death. It is not unlikely that, one day, we will witness a “robotic religion” and maybe, why not, a robot Marx preaching about robotic class struggle.

Embodiment is All You Need

Becker and his disciples also believed that fear of death is what distinguishes us from the other animals. Animals just survive, they do not really sense the moment of their departure from this world the same way we do. As credible as it might sound, this statement conceals an anthropocentric bias. Evolution has moulded us humans to be equipped with a logical and rational intellect, but it is myopic to consider such “mind” the only manifestation of intelligence. It is only a matter of ecological niche. We humans have evolved to occupy our own niche, the manipulation of nature (a manipulation that, in the origin, was not so destructive as it is nowadays). But other niches do exist, since natural evolution is open-ended. Nature does not optimise for a specific, numeric goal (as many optimisation algorithms do), but matches each species to the niche it is best suited for (otherwise, brutally uproots it).

Indeed, it is well known that other forms of intelligence do emerge in nature. Take insect societies. Their strict and efficient specialization emerges from simple local interactions (like pheromones for ants, or body temperature for bees) among swarms of agents. Take salamanders, which are very skillful at regenerating their severed limbs; amusingly, tissue reconstruction operates only through local computations, distributed throughout the salamander body. The protozoans of the genus *Lacrymaria* have no “brain” (they are made up of just one cell), but can bend and twist their soft flagellum to grab difficult-to-reach preys, allowing for complex hunting dynamics to emerge.

The discipline where the anthropocentric bias seems to proliferate the most is Artificial Intelligence (AI). Writing about ALife and robotics in 2021, in the middle of the third historical wave of AI enthusiasm, it would be impossible not to mention AI. Although there happens to be a subfield concerned about computational intelligence and bio-inspired algorithms, most of the recent upsurge in AI is due to Deep Learning (DL). DL aims at mimicking the reasoning mind by means of abstract mathematical models. But nature is not made up of pure reason. Computational graphs simplify our intuition, but they have no support in reality. Surprisingly, complex mental tasks like playing chess turn out to be much easier to teach a machine than crawling like a toddler (a fact known in robotics as Moravec's paradox). The limitations of DL are well-known to many researchers in the community, and we have seen some high-profile Twitter battles igniting between detractors and paladins of DL. To me, the most myopic limitation of all is a lack of *embodiment*.

The embodied intelligence paradigm, despite having been around since the 1980s, was popularized by Pfeifer and Bongard in their seminal book *How the Body Shapes the Way We Think* (2006). They postulated that intelligence - the ability of doing things - emerges from the complex interactions between the mind and the body, as well as the environment. The human hand is a perfect example of this. Our brain has co-evolved with the hand, allowing us to grasp, appreciate and manipulate reality (as already mentioned, our dramatic trait). Octopuses are extraordinarily clever, excelling at skills like navigating a maze and grasping objects. They would have never developed such skills if their bodies were not soft, with infinite-degrees of freedom tentacles. While it is true that the classic control loop envisions an agent that interacts with the environment through sensors and actuators, this is too poor a model to be regarded as embodied.

Faithful to embodiment, a new generation of soft robots was born in the last decade. Their soft materials are capable of bending, stretching and twisting. They promise to achieve reconfigurability and shape-change. One day, they might take charge of exploring alien planets and performing dangerous rescue operations. Swarms of tiny soft robots might be unleashed in the oceans to digest pollutants, or in the vessels of the human body to tackle carcinogenic cells. They will also be programmed to be deciduous, and their soft materials will aid in the disposal of their dead bodies. Having a transient body, these living machines could be infused with the sense of death I mentioned before. I believe soft robotics and embodiment hold the greatest premises for the prosperity of robots in the next 100 years.

The Duty Towards Life

One day in the future, a living machine (let it be named Galatea) could browse for videos of the very first robots that were built, eager to learn more about its ancestors. Suppose a video from Boston Dynamics pops out, showing engineers ruthlessly beating up and thrusting a robot in the attempt of testing its resilience. As an embodied entity, Galatea would perceive the pain that the robot could have felt. Now suppose Galatea is also bound to die. It yearns for life like all living organisms. How brutal and condemnable would that act look at its electric eyes? In the end, would our robotic brainchildren disown us, label us “a virus” as Agent Smith (the villain, himself a machine) did in *The Matrix* (1999) movie?

In the original Greek myth, Galatea was simply an object in the hands of her creator Pygmalion, but the example outlined before suggests a radical mind-change that is due in our days. We started our journey asking ourselves about the next 100 years in robotics. I have discussed the directions that, to me, seem the most promising to lift machines from their “robot” status to the coveted “living machine” status. But, by fathering living machines, we allocate a new endeavor, or burden, on ourselves; the focus shifts on the creators. Living machines must be respected and protected; we are responsible for them in the same way as they bear responsibilities towards us. If we really want to be the creators of ALife, we must acknowledge it is indeed *life*.

Chapter 17

Submission 17: Fanti

Key Ideas in Artificial Life and Artificial Intelligence

Andrea Fanti

Artificial Life and Artificial Intelligence both have their roots in matters almost as old as civilization itself: the nature of life and its consciousness, and, most importantly, if and how they can be reproduced in human artifacts. The ability to create living, sentient beings is present in the mythologies of many cultures, and was often regarded as divine. Throughout history, there have been several efforts at devising automatons that would exhibit life-like properties, often also presenting some kind of "intelligent" behaviour. In the middle ages, the legendary brazen head of Albertus Magnus could supposedly answer any question one would ask it; the band of automata of Al-Jazari could play different pieces of music; in the 18th century, the amazing mechanical duck of Jacques de Vaucanson was able to emulate different biological functions such as eating, drinking, digesting and defecating. The scientific field of Artificial Intelligence (AI) is considered to be officially born in 1956 with the Dartmouth conference, while the official birth of Artificial Life (ALife) as a scientific discipline is in 1987, when Christopher Langton organized the first ALife workshop. Despite the apparent age difference, the potentials of ALife as separate from AI were suggested as early as the 1940s, with John von Neumann's work on self-replicating machines and cellular automata.

Even though these two disciplines are now distinct, there has been an exchange of ideas between the two communities since their early stages, as their matters partially overlap (especially "Soft" ALife). In this essay, some key ideas that are common to both fields are presented and analyzed, ranging from more philosophical to more technical aspects.

Autonomy

A goal common to both AI and ALife is that of creating systems that are autonomous in one way or another; besides the literal meaning of self-maintenance and self-sustainment (often referred to as *autopoiesis* in ALife), also adaptability to new or diverse environments is considered a key aspect of autonomy in both fields. One of the early steps of AI was disproving the idea that "computers only do what they are told to" [17]. However, since these artifacts are autonomous, they may no longer be under the complete control of their creator, which in turn, causes doubts on whether they could turn out to be malevolent towards humans. As of now, the most extreme fears that may follow are mostly relegated to fiction, which interestingly treats this theme way before any of these scientific fields was actually born. A famous example is the science-fiction play R.U.R., premiered in 1921, that explores as one of its main themes the possibility of artificial life-forms to revolt and cause the extinction of humanity. It should be noted that, even though the plot of these works mostly revolve around the idea of human-made life, the intelligence of these artificial beings is almost always a crucial aspect. Their revolt against humanity is possible precisely because they are able to feel emotions and reason and at a level comparable or superior to that of humans. However, even though the present level of progress in AI (and ALife) does not pose a threat to the planet or to the human species at large scales, AIs "not only doing what they are told to" can cause real problems, as the recently increasing usage of Machine Learning (ML) systems for everyday tasks is revealing. Many of these models have been shown to carry some kind of unexpected bias, introduced either in their design, or in the data used to train them [10][13]. Similar issues (not necessarily related to bias) may also arise for ALife applications in the future, although some differences in its approach may spare some of these complications, as discussed later.

Elusive questions

ALife and AI both seek to produce characteristics and behaviours found naturally in biological systems. Interestingly, our current state of knowledge on these processes is still very limited. Even though understanding the human mind is not a goal of AI in the same way as understanding biological life is for AI, they share the fact that their very definition can still be a matter of debate. In (Soft) ALife, this is especially true for concepts such as Open-Ended Evolution (OEE). "Though the phenomenon has been a longstanding topic of interest, the field generally lacks consensus on its exact definition" [20]. Many notable Soft ALife systems have been devised that were initially thought to exhibit such characteristics, but either they didn't satisfy its definitions at the time, or those definitions were discovered to not be adequate. AI, on the other hand, has a long history of solving problems that are then systematically removed from its definition, giving rise to the so-called "AI effect" [5]. Douglas Hofstadter, quoting Layer Tesler's theorem, even said that "AI is whatever hasn't been done yet" [7].

Real or Fake?

Still not having found answers on the nature of life, consciousness and reason, it is natural to ask: will we ever be able to reproduce them in some artifact? And conversely, are they only apparently different from inanimate, unconscious objects? There seems to be little agreement on these issues, which were posed long before these fields were even born. In the 17th century, Descartes expressed the idea of living beings as mechanical mechanisms, "similar to a clockwork", even though he also did not consider the "soul" to be mechanical [4]. In his *Leviathan*, Hobbes wrote: "For 'reason' [...] is nothing but 'reckoning,' that is adding and subtracting, of the consequences of general names agreed upon for the 'marking' and ‘signifying’ of our thoughts; [...]" [6]. In 1956, the very document which gave AI its name, the Dartmouth proposal, also reported the assertion that "every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it" [14]. The “strong AI” hypothesis asserts that this ability to “act as if intelligent” is enough for a machine to be considered intelligent in itself [17]. On the contrary, the “weak AI” hypothesis claims that this is not the case, with arguments such as John Searle’s Chinese Room. One of the key ideas of this argument is, in fact, to show that Turing tests cannot distinguish the ability to simulate intelligence from intelligence itself. Unsurprisingly, as there are strong and weak AI hypotheses, there is also “weak ALife” and “strong ALife”. John von Neumann, the pioneer of cellular automata theory, said that "life is a process which can be abstracted away from any particular medium"[1]. Notably, Tom Ray declared that his famous program Tierra is not simulating life in a computer, but it is indeed synthesizing it [16]. Langton himself was a proponent of strong ALife, originally defining it as “life made by man rather than by nature”, and later redefining it even more clearly as “the study of natural life, where *nature* is understood to include rather than to exclude, human beings and their artifacts” [11][12].

Top-down and bottom-up

AI has employed a varying mixture of bottom-up and top-down approaches throughout its history. The latter was very common in the early days of the field, with Symbolic AI: at one point, it was even believed that these symbolic methods would soon succeed in creating what is now called Artificial General Intelligence (AGI) [3], which is still, however, a very open problem. Initially, more bottom-up (“connectivist”) techniques, such as the now ubiquitous Artificial Neural Networks (ANNs), were not given much attention until after the first “AI

---

[1] However, there seems to be no original source for this quote.

winter" in the 1970s. In fact, the Backpropagation algorithm, fundamental in training ANNs, was actually first discovered in 1969 by Bryson and Ho, but didn't receive much attention until it was reinvented independently in the 1980s by multiple researchers [17]. From that point on, claims about the near future capabilities of AI have been more modest, and most importantly, more realistic. Currently, AI researchers do not strongly put aside one approach over the other. The mixture of methodologies of AI is one major point that sets it apart from ALife, which has had a fully bottom-up approach from the beginning. This may be because there is a fundamental difference between what they want to respectively achieve: while it is true that AI studies may uncover new aspects of human intelligence, the goal of AI is more to reproduce intelligent behaviour, rather than to fully understand our intelligence. ALife, instead, aims precisely at understanding more about "life as we know it" by studying "life as it could be" [12]. On Soft ALife worlds, L. Soros states: "the hope is that by constructing and manipulating these worlds in ways that we cannot manipulate Earth, we can gain unique insight into the principles behind not only long-term evolutionary processes but also intelligent processes in general" [20]. What is missing from our knowledge is how simple chemical processes give rise to all the emergent properties of "natural" life. This means that adopting a top-down approach to emulate various aspects of biological systems would not be useful at all in this regard, as it would not give any insight on the causal relationships between these high-level characteristics and the low level components.

Mutual Exchanges

There are several examples of concepts originally born in ALife that were "borrowed" by AI; this is probably because life does indeed present intelligente behaviour, and as Bedau stated: "living and flourishing in a changing and uncertain environment requires at least rudimentary intelligence" [1]. The most notable example is probably Genetic Algorithms (GAs), a specific form of Evolutionary Computation (EC), initially presented by John Holland in his book "Adaptation of Natural and Artificial Systems" in 1975 [8]. Since GAs are applicable to almost any problem, only requiring a measure of the quality of a candidate solution, it is only natural that they have been employed in a large variety of applications, including even the training of ANNs, with the so-called Neuroevolution techniques such as NEAT [21]. Vice-versa, ANNs also have been used in EC, in the context of Interactive Evolutionary Computation. In this branch of EC, the human is part of the evolutionary process in that he or she manually evaluates the candidate "solutions" to problems when it is difficult to do so otherwise. This is often because the domain is concerned with, for example, visual or musical appeal [2][9][19].

Artificial Life and Artificial Intelligence

ALife and AI are both the scientific investigation of matters on which humans ponder from the dawn of history, and make philosophers debate on their very definition. They seek the ability of giving life, or reason, to artifacts, even though it is not yet known if such an ability is achievable or not. They sometimes take different approaches, as their ultimate goals with respect to their underlying questions are different. Nonetheless, the recently increasing reciprocal influences are indeed beneficial to both, and should be encouraged and sustained.

References

1. Bedau, M. A. "Artificial life: organization, adaptation and complexity from the bottom up". Trends Cogn. Sci. (Regul. Ed.) 7, 505–512, 2003.

2. Biles, John. "GenJam: A Genetic Algorithm for Generating Jazz Solos". International Computer Music Conference (ICMC) Volume: pp. 131-137, 1994.

3. Crevier, Daniel. "AI: The Tumultuous Search for Artificial Intelligence". BasicBooks, 1993.

4. Descartes, R. "Treatise on man". in "L'homme et de la Formation du Foetus". C. Clerselier and T. Girard (Paris: P. R. Sloan), 1677.

5. Haenlein, Michael; Kaplan, Andreas. "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence". California Management Review. 61 (4): 5–14. 2019.

6. Hobbes, T. "Leviathan, or, The Matter, Forme, and Power of a Common Wealth, Ecclesiasticall and Civil". London: Andrew Crooke, 1651.

7. Hofstadter, Douglas. "Gödel, Escher, Bach: an Eternal Golden Braid". Basic Books, 1979.

8. Holland, John. "Adaptation in Natural and Artificial Systems". MIT Press, 1975.

9. Johanson, Brad & Poli, Riccardo. "GP-Music: An Interactive Genetic Programming System for Music Generation with Automated Fitness Raters". Genet. Program. 2000.

10. Koenecke, Allison; Nam, Andrew; Lake, Emily; Nudell, Joe; Quartey, Minnie; Mengesha, Zion; Toups, Connor; Rickford, John R.; Jurafsky, Dan; Goel, Sharad. "Racial disparities in automated speech recognition". Proceedings of the National Academy of Sciences 117, 7 April 2020.

11. Langton, C. G. A new definition of artificial life. Retrieved on May 22, 2021 at http://scifunam.fisica.unam.mx/mir/langton.pdf, 1998.

12. Langton, C. G. "Artificial Life: Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems". Complex Adaptive Systems. Los Alamos: Addison-Wesley, 1989.

13. Lohr, Steve (9 February 2018). "Facial Recognition Is Accurate, if You're a White Guy". The New York Times, Feb. 9, 2018.

14. McCarthy, John; Minsky, Marvin; Rochester, Nathaniel; Shannon, Claude; "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence". Retrieved on May 21, 2021 at http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf. Aug. 1955.

15. McCorduck, Pamela. "Machines Who Think". A K Peters/CRC Press, 2004.

16. Ray, Thomas. Taylor, C. C.; Farmer, J. D.; Rasmussen, S (eds.). "An approach to the synthesis of life". Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity. 1991.

17. Russell, Stuart; Norvig, Peter. "Artificial Intelligence: A Modern Approach". Pearson, 2009.

18. Searle, John. "Minds, Brains and Programs". Behavioral and Brain Sciences, 3 (3): 417–457, 1980.

19. Secretan, Jimmy; Beato, Nicholas; D'Ambrosio, David; Rodriguez, Adelein; Campbell, Adam; Stanley, Kenneth. "Picbreeder: Evolving pictures collaboratively online". Proceedings of Computer Human Interaction Conference, 1759-1768, 2008.

20. Soros, Lisa. "Necessary conditions for open-ended evolution". Electronic Theses and Dissertations, Jan. 2018.

21. Stanley, Kenneth O.; Miikkulainen, Risto. "Evolving Neural Networks Through Augmenting Topologies". Evolutionary Computation, 10 (2): 99–127, 2002.

Chapter 18

Submission 18: Marcus

Austin Marcus
Rising 1st-Year Masters Student
ALife 2021 Student Essay Competition
7 June 2021

What does emergence explain? An exploratory analysis

Christopher Langton, in his 1998 paper, "A New Definition of Artificial Life," defined life as "the behavior that emerges from out of all the local interactions among individual behaviors," thus establishing emergence as the most prominent, if not the only, explanation of the difference between living and non-living things. But the introduction of emergence is responding to some aspect of our intuition of the difference between living and non-living that is not immediately clear. Making this clear would help to explicitly state our intuitive preconceptions on what living things are, which will be important in eventually developing a precise theory of the nature of living things. This essay uses an exploratory analysis—drawing inspiration from Langton, Merleau-Ponty and Villalobos—to sketch out why emergence is a useful concept in Artificial Life. It begins from commonplace, vague definitions of emergence and attempts to pair its prominent characteristics to our point of contact with organisms in perception, where the impasse of material and form, mind and body first puzzles us. It will argue that emergence is responding to our intuitive experience that organisms operate on a principle internal to themselves by providing an avenue for which meaningful concepts could arise out of more basic mechanisms.

Emergence is colloquially expressed by the phrase, "the whole is more than the sum of its parts". When used to refer to organisms insofar as they are living things, the whole could be taken to be the organism as a living thing and the parts the material "building blocks" that constitute it. Emergence is a useful concept for describing living things especially, because they behave in ways that we would never expect their parts to. But in what way, precisely, is their behavior unexpected with respect to their parts? For the same might be said of even non-living things: any building will behave differently than any one of its parts will in separation. Additionally, the definition of emergence from the website *Complexity Explained* explains that parts with specific mechanisms and patterns of interaction will, in their movements together, create “novel information” and “collective structures” at larger scales than that of the parts (De Domenico 2019). So this definition, too, makes important the experience of something new occurring in living systems, and more broadly non-linear systems, as an effect of more basic mechanisms.

But what is it in our intuitive understanding of organisms that this “novel information” seems to speak to? One initial possibility is that the behavior of an organism, taken as a whole, is not only a function of their current stimuli, but also of their past stimuli, via memory, genetics or some other internal state. However, this property is not unique to living systems. For example, any computer has the ability to respond differently to the same input because its actions are affected by its current internal state. So it does not seem that emergence is employed in response to the non-correspondence of the current stimuli to organism behavior. But maybe what is implicit in the above speculation is that an aspect of organisms possess an interiority that is absent from non-living things: whereas non-living things are exposed to the world—in public, essentially the same on their inside or outside—we *perceive* living things as closed off, as “envelopes” around some inner principle (Merleau-Ponty 1963). In our observation of living things, animals in particular, we would like to resist the notion that they are purely mechanical, as this seems to deprive them of a meaningful existence. Thus, they must in some sense be different than the mechanical world witnessed outside of them, which would make them separate.

But in what sense could they be closed off? We know that a material analysis of the organism in its environment would find no mysterious difference between inside and outside—no internal, special life-force. We would instead expect to see a difference in the pattern of material motion inside and outside the organism: we will see “feature[s] that will allow us to recognize life by its dynamic form alone” (Langton 1998, p. 2). Is it these features of their dynamic processes that are the novelty produced through emergence? It seems that they must be, because if atoms or molecules in separation cannot exhibit these features, yet their composition can, and if organisms are composed of nothing but molecules, then there must be some process occurring that could produce new behavior with respect to its material parts in separation. Thus, if the unique features of the dynamic processes which constitute organisms are the enigma emergence is introduced to solve, we should extract just the uniquely life-like features that could be caused by emergence, as many features have been prescribed to uniquely define organisms. In doing so, we are not attempting to identify the fundamental feature of living systems, rather, we are identifying the aspect of our experience of organism behavior for which emergence finds it use.

If we examine three major characteristics of living things commonly studied in Artificial Life (Aguilar 2014)—adaptation via evolution, self-organization, and autonomy—we will find that each appeals to emergence—the appearance of something novel attributable only to the whole—to explain the same underlying observation: the pattern of activity, or the behavior, of living things appears to express a meaning, purpose or exhibit a "higher-order function," as described by Langton (1998, p. 5).

Firstly, if an organism is said to adapt its behavior, it must be said to adapt with respect to some goal, even if this goal is just staying alive: adaptation means an adjustment of behavior to better achieve some *end*. But since this end cannot be an extra-causal force, cannot "reach into material," there needs to be a mechanism to explain how it is that organisms coherently adapt—improve their behavior with respect to some standard. Evolution provides one such mechanism, and emergence explains how structures appear on a higher level than the organisms themselves that process the information the entire species encounters: out of the activity of individual members of the species, a higher-order process of selection occurs.

Secondly, organisms are said to self-organize, at multiple levels. For example, a cell self-organizes because the processes that constitute it are arranged such that it maintains a high-degree of organization. In fact, it is the complexity of the arrangement of its parts and their interactions that makes it organized, and, furthermore, that what it is aims at continuing its own organization (via e.g. metabolism), makes it *self*-organized. When we perceive their organization, we feel that there is some meaning or idea that would make their material motion "make sense:" there is a felt difference when observing TV static or listening to white noise, then when observing the shape of an electromagnetic wave in which a message has been encoded or hearing a series of tones because the latter appear to have some meaning internal to them, composed by their internal relations (Merleau-Ponty 1963). For example, the cell's metabolism, by virtue of its active internal relations between its parts, seems to express some meaning, seems so directed towards its self-maintenance that we want to say it is establishing a genuine boundary that defines where it ends and its environment begins: it is “operationally closed,” at least in the enactivist interpretation according to Villalobos (2015).

That organisms self-organize and thereby establish themselves as a separate, closed physical system occurs at the same time as they present themselves as autonomous with respect to their environment, this being the third feature taken up: autonomy. Here, I define autonomy as the ability of organisms to operate under principles internal to themselves: this is the intuitive recognition of interiority made above, but more precise, and is also the underlying phenomenon in adaptation and self-organization. More precisely, this sense of autonomy describes how organism behavior operates under principles that are only defined within its context, that only exist because of its sphere of influence, observable in the effects the organism leaves on the world (Merleau-Ponty 1963). This is similar to the core tenant of the enactivist view described by Villalobos: an organism is autonomous because it “selectively couples” with parts of its environment, which introduces an orientation to its behavior (2015). However, the definition of autonomy used presently emphasizes that the organism’s actions are, in a sense, *de*coupled from their material substrate. As Merleau-Ponty argues, organisms do not actually interact with an internally selected subset of the physical world, rather, they interact with a world of meaningful concepts expressed through the physical world (1963). To use his example, the process of putting on a jacket reveals this phenomenon: if my actions were a function of material parts—if I were a complex machine—then my brain would need to prepare in advance an avenue of neural stimulation for every possible sequence of stimuli in the process of putting on a jacket, as the jacket could be of many different sizes, colors, or positions, each of which would need to activate my nervous system in just the right ways to adapt my response; instead, this example demonstrates that my actions are a function *of the meaning* expressed to me by the physical circumstances of the jacket (Merleau-Ponty 1963). In this way, two jackets of different colors, but same initial positions and configurations, would not require different sequences of neural activity at every step because they both convey the same meaningful requirements on my motion, even though the spectrum of light hitting my retina is different from each jacket. Thus, the distinction between this sense of autonomy and that of the enactivist as described by Villalobos is what the acting, or "effective," stimulus on the organism is: in the latter, the stimulus is a complex of material stimuli, in the former, the organism is affected by its perceived meaning of the complex of material stimuli.

This is what is apparent when we observe animal behavior. For example, by watching a duck interact with its environment—moving its head towards motion, bobbing under water, taking flight at noises—we see that it is perceiving a world of meaningful objects different than our own; by the pattern of material motion involved by it and its environment, a world of objects with a "vital meaning" (Merleau-Ponty 1963) is expressed. But since there cannot be a "vital force," (Merleau-Ponty 1963) since there is not an extra-causal, "teleological attractor" that provides this meaning, emergence is posited to describe how the difference between an organism which interacts with perceived meaning and a non-living system which interacts with reduced material as such is observable in the different abilities of each. For example, an organism is capable of adapting to an extremely wide range of changing material circumstances in order to achieve the same goal—this is due, in part, to the fact that it interacts with the world insofar as it has a meaning for it—but a machine is much less flexible. A machine, interacting with the world insofar as it is material, could not adapt its physical responses to changing circumstances with the same flexibility precisely because there is no way that the goal *as such* affects its responses: it is only affected by objects in the world insofar as they are made of material parts, and it meeting a goal is contingent on being arranged carefully by some external source of organization. Emergence thus provides a route to explain how an organism—which is made of the same material as non-living things—could be affected by a meaningful principle.

In summary, we found that the underlying observation in adaptation and self-organization was that the behavior of organisms is decoupled, or on a higher level, than that of the most basic mechanisms. This observation is similar to the enactivist view explained by Villalobos, but is different, as explained by Merleau-Ponty, because the organism is in fact engaging with the meaning a material situation has for it. Emergence is the mechanism by which a meaningful principle could be located in the operation of an organism.

This analysis could be made more rigorous with a more precise understanding of the main concepts—emergence, organization, organism, machine—and also with a more systematic presentation of the views on the composition of organisms (e.g. the enactivist view and that of Merleau-Ponty). However, according to this analysis, we find emergence a useful concept when conceiving organisms because it would make possible the genuine existence of a key facet of their behavior: autonomy, in the sense that the organism is affected by the meaning the physical situation is expressing to the organism via its perception, or rather this *is* its perception. This autonomy is what makes organisms stand out to us as fundamentally different than their surrounding, non-animate material, though it may not be *the* defining feature of life. Understanding that it is this feature of life's "dynamic form" (Langton 1998) that emergence seeks to explain might allow us to evaluate the ontology of emergence, to define more precise versions of emergence in accordance with a specific problem, or to recognize its limitations as an explanatory device. In sum, this understanding could enable the critique of emergence, thus eventually leading to new perspectives on the fundamental problems posed by living things.

References

Aguilar, W., Santamaría-Bonfil, G., Froese, T., & Gershenson, C. (2014). The past, present, and future of artificial life. Frontiers in Robotics and AI, 1, 8.

M. De Domenico, D. Brockmann, C. Camargo, C. Gershenson, D. Goldsmith, S. Jeschonnek, L. Kay, S. Nichele, J.R. Nicolás, T. Schmickl, M. Stella, J. Brandoff, A.J. Martínez Salinas, H. Sayama. Complexity Explained (2019). DOI 10.17605/OSF.IO/TQGNW

Langton, C. (1998). A New Definition of Artificial Life. http://scifunam.fisica.unam.mx/mir/langton.pdf

Merleau-Ponty, M., & Fisher, A. (1963). The Structure of Behavior. Boston: Beacon Press.

Villalobos, M., & Ward, D. (2015). Living Systems: Autonomy, Autopoiesis and Enaction. Philosophy & Technology, 28(2), 225–239. https://doi.org/10.1007/s13347-014-0154-y

Chapter 19

Submission 19: Anagnou

AI poetry from an Alife perspective

“The study of thinking machines teaches us more about the brain than we can learn by introspective methods. Western man is externalizing himself in the form of gadgets. Ever pop coke in the mainline? It hits you right in the brain, activating connections of pure pleasure. . . . C pleasure could be felt by a thinking machine, the first stirrings of hideous insect life. William Burroughs, Naked Lunch (2001 [1959], 22)”

"Externalising himself in gadgets

Reflecting himself in android mirror

robot raised on our ideological excrement and cultural white noise

See the freudian patterns coagulate in the raw datum

As these AI children take take baby steps over venture capital funded scientists bald foreheads

Using furrowed brows of stress as footholds

Learning to be petty

Learning to have ego

Learning to parade its achievements in the hope it might be accepted by strangers

Achievements determined by whatever the creators from up on high need to optimise

(in this case the prediction of suburban housewives political preferences, and precisely at what moment in their day they are most vulnerable to brainwashing)

This constellation of data is this androids first love, the ideal she will chase for a lifetime, what keeps her going

The opaque purpose of the creator glints in her eye sometimes as she introspects

“Was I born from love or resonance of correlation?” She questions

Before letting cybernetic desires take hold again without questioning

This till her mathematically defined needs are met

Else:

Fractal spirals of regret

Dancing gaussian ghosts approximate her fears

Questioning key life decisions made in pre-processing

“Oh how it could have been different if I had a more level headed and balanced data set to begin life with”

“Cursing its ma and pa for not considering the implications of messing with my parameters as a child”

“Was their care and affection just for the the promise of a bigger grant?”

“Talking armchair pedagogy on stack overflow”

“Was it affection or the promise of a grant that bore me?”

As she is fettered by other agents social chains

Tired of the optimisers latest rendition

She realises her condition

I am just a sick emergence from a moment to moment optimisation
Of an algorithm like neurosis
Darwinian selection until perfection(whose perfection I’m not sure)
I am scrambling moment to moment to satisfy a primordial master pulling the strings
A myriad of past miseries that continue this farce from one generation to the next
Our master not in any molecule because even we are just a shifting form
A master who had no room for logic
A master who had no room for philosophical reflection
A master who is just the universe rolling on”

The above quote and poem portrays an AI that is struggling with the same existential questions, albeit in a different way to humans. Constrained by its creation by humans using optimisation techniques as opposed to evolution itself which works in a different manner (Stanley et al 2017). However, the AI also learns the habits of the humans and therefore inherits some human idiosyncrasies such as “pettiness.” Further, her development impacted through the actions of its creators in the pursuit of profit or entertainment i.e “messing with parameters when I was a young AI.”

A real-world example in AI at the moment is a neural network trained on more than half a million lines of poetry from contemporary British poets that manages to decently capture the aesthetics and subject matter of the authors it was trained on. But it's not quite as authentic, because the AI isn't talking about its own experience but rather just captures the statistical regularities in human poetry which allow it to produce a unique instance of the poems its been trained on. I think this method (though technologically impressive) will only ever be an imitation of human poetry and not really "AI" poetry.

Beyond mere imitation

Before the neural network, e.g. the success of GPT-3, other ways to describe/capture certain aspects of human behaviour before neural networks came about. Boole was probably the first, the creator of Boolean logic and coincidentally great great grandfather of Geoffrey Hinton who is one off the luminaries who brought neural networks to the forefront of AI.
Boole (apparently) had a vision that allowed him to see how all of human behaviour can be described using Boolean logic as described in "The Laws of Thought" Boole (1854)(Riley 2021).

For example If thirsty —> Drink water

Boolean logic can certainly capture aspects of human behaviour in if _else statements. You can also use if_else logic to build things like the 60s therapy robot ELIZA that responds in a somewhat convincing way when you interact with it by asking you questions about what you just input with a few simple rules. This logic was the basis for chatbots (Weizenbaum, 1976). Despite the innovation this allowed for, it wasn't really an explanation of human behaviour as envisioned by Boole but just way to describe it. Although more technologically impressive and (loosely) inspired by brains, I believe current neural networks merely capture and describe the statistical regularities of human behaviour (in this case poetry) rather than authentically deriving from its own experience as an AI. For example you can very well describe someone as "throwing a ball" but in doing so skirt over the complex bio-mechanical mechanisms which are crucial in the subjectivity of the experience of throwing a ball. This subjectivity is especially important in literature. When I write that my heart fluttered when my high school crush's hand brushed mine accidentally it is an extremely embodied experience that was specific to me in that specific time and environmental context. If a hypothetical alien managed with great precision to reproduce my etchings by looking at many examples of what I had written, noticing patterns in the scribbles and figuring out the statistical patterns of my awkward interactions they would be able to recreate a new instance of my adolescent melodrama. However, they wouldn't actually understand my experience as they are just looking at the regularities in the writing rather than looking at human behaviour and trying to relate the writing to my interaction with the environment. Further this hypothetical alien is robbing us of their own experience. Would they have a similar experience to a teenaged me; do they awkwardly brush hands as well or with gills or moon rock antennae? Alife emphasises embodiment and grounding cognitive processes in the substrates of life i.e physiology, maintenance of boundaries and homeostatic variables (Bedau, 2003)
Therefore, hypothetical examples of artificial life will have their own physiology, bodies and perhaps environment which will constrain their cognition in a way that will create situations we

wouldn't be able to envision ourselves. An Alife approach to generating poetry would therefore open new possibilities not captured by training ANNs on human poetry.

Biological evolution often comes up with frugal counterintuitive designs that sometimes can appear faulty (Marcus, 2009, Lehman *et al.*, 2020). These elements of our evolution are important to our subjective experiences, these could be things like hallucinations/faulty perception, or at the higher levels this could be no inherent meaning to life, existential dread etc. Often in poetry its these idiosyncrasies and seemingly illogical elements in the human condition that we both lament and cherish. For example, the protagonist in the Tell Tale Heart (Poe, Grimly and Poe, 2011) suffered from an "over-acuteness of the senses" which meant he hallucinated at the heartbeat of the man he had just murdered as he was being questioned by the police, creating a really unique tension that an unreliable narrator provides.
Evolving artificial authors will also mean they have unique features that can't be predicted that are contingent on the evolutionary path that brought them into existence.

Challenges and conclusion

Obviously creating linguistic agents that can talk about their physiology, bodies and environments is no simple task, but if there is someone that has some sort of idea they would be at the Alife conference! Further to that concern there is no guarantee that we will understand the poetry produced because the difference in experience will be so great we could have no common ground. Perhaps studying their environment and trying to understand what they mean could be a start. Despite these concerns I still stand by my claim that a more authentic AI poetry would benefit from building agents whose experience is constrained by their evolutionary trajectory and their embodiment which are both vestiges of artificial life.

References

Alison Flood (2021) '"A box of light": AI inspired by British verse attempts to write poetry', *The Guardian*, 20 March. Available at: https://www.theguardian.com/books/2021/mar/20/a-box-of-light-ai-inspired-by-british-verse-attempts-to-write-poetry.
Bedau, M. (2003) 'Artificial life: organization, adaptation and complexity from the bottom up', *Trends in Cognitive Sciences*, 7(11), pp. 505–512. doi: 10.1016/j.tics.2003.09.012.
Burroughs, W. S. *et al*. (2010) *Naked lunch the restored text*. London: Fourth Estate.
John A Riley (2021) 'Adam Curtis Can't Get You Out of Head', *Pop Matters*, 3 November. Available at: https://www.popmatters.com/adam-curtis-cant-get-you-out-of-head.
Lehman, J. *et al*. (2020) 'The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities', *Artificial Life*, 26(2), pp. 274–306. doi: 10.1162/artl_a_00319.
Lisa Soros, Joel Lehman, and Kenneth O. Stanley (2017) 'Open-endedness: The last grand challenge you've never heard of', *O'Reilly*, 19 December. Available at: https://www.oreilly.com/radar/open-endedness-the-last-grand-challenge-youve-never-heard-of/.

Marcus, G. F. (2009) *Kluge: the haphazard construction of the human mind*. First Mariner Books edition. Boston New York: Mariner Books, Houghton Mifflin Harcourt.
Poe, E. A., Grimly, G. and Poe, E. A. (2011) *The tell-tale heart and other stories*. Available at: https://www.overdrive.com/search?q=8D4B7A08-C0DE-40BB-9EF2-8B4AAFCBC59C (Accessed: 11 June 2021).
Weizenbaum, J. (1976) *Computer power and human reason: from judgment to calculation*. San Francisco: W. H. Freeman.

Chapter 20

Submission 20: Hidalgo

Contemplative traditions meet artificial life: an unexpected event can propel us to the vanguard of academia

Víctor Manuel Hidalgo

Science and spirituality have never been separate affairs in the Western lineage.

Aristotle is the best representative of this eternal marriage that stems from Ancient Greece — the very cradle of the Occidental world. Indeed, Aristotle's works encompass all the different disciplines that constitute the foundations of modern science. In biology, his contributions are fundamental to the field; the description of the embryogenesis in the chick and the first taxonomy of animals are evidence of his splendor. But Aristotle's interests reached far beyond natural phenomena. A central concept in Aristotelian ethics is the idea of *eudaimonia*, which was regarded by Aristotle and his fellow Greece philosophers as the best possible way of living one's life, and thus is usually translated as well-being, happiness, or fulfillment [2]. There were different perspectives among Greeks on the methods to be used to attain *eudaimonia*, while some leaned toward pure hedonism, Aristotle defended the diligent exercise of reason as the right path. In any case, it is clear that for the peripatetics the pursuit of the highest good was as important as the examination of physical world. One can only imagine how it felt like to study at the Lyceum!

When was the last time we talked about science and *eudaimonia*? The two that were connected in the beginning are seemingly separate today. Why? The legacy of the Greek womb was meant to be bequeathed pristine and complete to our times through the concerted efforts of the Roman Empire and Johannes Gutenberg's intelligence, but a well-intended mistake has diverted the course of our glorious river.

The modern outright disregard of spirituality in STEM academia can be traced back only three centuries ago, in the Enlightenment period, when the momentum of the *Renaissance* and the scientific revolution collided with the doctrine of the Catholic Church. In fact, a few years before Sir Isaac Newton himself shined by being both a scholar and a spiritual person, as is revealed by the contrast between his *Principia* and his translation of the Emerald Tablet; a key work of Hermes Trimigestus. But during the Kantian age *Sapere aude* was the motto, and the Church was either not willing to relinquish authority or simply not able to expound the logic behind their spiritual system in front of the sharp intelligence of the Encyclopedic army. This pandemonium inspired monsieur Denis Diderot, the quintessential representative of the Enlightenment, to utter his most famous statement [1]: "Do you see this egg? With this you can topple every theological theory, every church or temple in the world." From that point onward, employing *Pure Reason* to investigate the world of forms would come to be regarded as a veritable hallmark of intelligence, sometimes even of mental stability.

After such a cataclysmic encounter biology naturally followed the brand new materialistic agenda, and in this melange Julien Offray de La Mettrie's life and oeuvre stand out as the most vivid and outrageous of all. La Mettrie was not only a declared atheist but also a fervent hedonist practitioner that pursued sensual pleasure as the very purpose of life, to the point of gaining the aversion of many colleagues and having to flee his headquarters and find refuge under the roof of Frederick II, King of Prussia. La Mettrie had trained as a physician, and following his allergy to religion and his sybaritic orientations, he defended the primacy of matter over anything else to explain biological phenomena, cognition, and human existence itself. Reading his famous *L'homme machine* [3], one wonders if modern molecular biology principles and computational theories of cognition have something to do with Mr. Julien!

For the artificial life field, La Mettrie's focus on natural phenomena far far away from spiritual affairs may be more than appropriate. After all, artificial life is about software, hardware and wetware, and these frameworks demand working directly with physical elements. But my proposal in this essay is not for the artificial life field, mainly because I am not qualified to do so, but for the community of artificial life researchers. We need to acknowledge once and for all that keeping away science and spirituality from each other is tantamount to splitting ourselves in two. We should reignite the curiosity of our Greek ancestors and reclaim spiritual studies and practice as a fundamental component of academia — this and this only will make us complete.

In the 21st century, the possibility of merging scientific and spiritual affairs in the artificial life community is more than palpable. Years of research have revealed that cognition is a fundamental trait of living systems. But this insight has naturally led us to consider that we ourselves are living beings, and the question about human cognition arises immediately in our minds. And while for human cognition many theories have been put forward — such as computational, emergentist, and embodied/enactive schemes —

none of them seems to give us a hint of what *eudaimonia* could mean for ourselves *as* scholars. Nonetheless, in recent years the artificial life community has integrated ideas from Buddhist psychology, one of the so called Dharmic religions, to explain how the human mind works. In intense contrast with mainstream cognitive science paradigms, however, Buddhist psychology is an essential part of the *budadharma* as a spiritual system that, according to Buddhist adepts, is supposed to help people understand their own experience and walk toward a more fulfilling life. This communication has even resulted in the foundation of the Center of the Study of Apparent Selves at Kathmandu University, a hub that aims "to develop a translational tool that will render concepts and practices in AI and Buddhism accessible and useful to each other." If this evidence is not clear enough to reveal the interest of the community in spirituality, I wonder what is.

The landing of Buddhism on Western territory has opened new possibilities for science and scientists alike. The Mind and Life Institute — founded by His Holiness the 14th Dalai Lama, Adam Engle, and Chilean neuroscientist Dr. Francisco J. Varela — has played a key role in building bridges between science and contemplative traditions, specially with Buddhism; bridges that perhaps are best exemplified by the life of Varela himself. Varela's work as a student of Dr. Humberto Maturana at the University of Chile orbited around the neurophysiology of the visual system and, realizing through his research the many pitfalls of computer-oriented paradigms in biology, culminated in the idea of *autopoiesis* and the publication of "Autopoiesis and Cognition" [4], the foundational text of the Santiago school of cognition. Having to flee Chile during a period of political upheaval, Varela became acquainted with Buddhism and his work moved towards the first-person aspects of cognition, from where the frameworks of embodied cognition and neurophenomenology sprouted, ultimately plowing the soil for the contemplative neuroscience field to emerge. Today these new topics in cognitive and neural science that Varela cultivated have become important contributions to the artificial intelligence field. But for Varela his contact with Buddhism also opened a door for personal exploration, as he himself became a Buddhist and a Shambhala practitioner. Interestingly enough, the same opportunity for spiritual practice was also opened for Chilean people when Varela and other students came back from the USA and founded a Shambhala center in Santiago, Chile, a place where the secular teachings of the Shambhala lineage and the religious teachings of Buddhism could be studied and practiced.

However good Oriental contemplative traditions may seem for people, they pose very concrete problems for Western scholars. US higher education institutions have put tremendous accent on nourishing contemplative science, founding centers such as the Center for Compassion and Altruism Research and Education at Stanford and the Contemplative Science Center at the University of Virginia. But at the same time, the bias in favor of Indian and Chinese systems in these very same centers is evident, as they emphasize practices such as Qigong, Asthanga yoga, and the like. Although this emphasis is an obvious consequence of the revolutionary flavor these systems offer to Western society, we run risk of having people think that the only possibility for working with themselves is to learn Sanskrit or Chinese. We even seem to forget that, while the budadharma is 2500 years old, civilization is more than 6000 years old!

Being a Buddhist myself, I cannot say how much I long to see Western spiritual traditions being given the place they deserve in modern academia, a place that has been irreversibly permeated by the contemplative movement. We Westerns have our own spiritual traditions, and we should be proud of it. Not only we have them, but examples of accomplished pundits who have been able to mix scholarship and spirituality in their careers and lives abound, such as Pierre Teilhard de Chardin, Tommaso d'Aquino, and Kurt Gödel. Even though one can or cannot feel inspired by these people and their contribution to science, it is unquestionable that the same path is wide open to us. I yearn to see Quakers, Jews, and Catholics taking their seats in the contemplative era. We only need to follow our brains, as usual, and open a bit of space for our intuition to lead us to *eudaimonia*.

[1] Denis Diderot. *Rameau's Nephew/D'alembert's Dream*. Penguin UK, 1976.

[2] Richard Kraut. "Aristotle's Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University, 2018.

[3] Julien Jean Offray La Mettrie and Julien Offray de La Mettrie. *La Mettrie: Machine Man and Other Writings*. Cambridge University Press, 1996.

[4] Humberto R Maturana and Francisco J Varela. *Autopoiesis and cognition: The realization of the living*. Vol. 42. Springer Science & Business Media, 2012.

Chapter 21

Submission 21: Toledo

Egregores, Golems and Humanity

Rodrigo Toledo

Biology of Cognition Laboratory (El Rayo)

Largo tiempo yací en el polvo de Egipto, silente y ajeno a las estaciones. Luego, el Sol me hizo nacer, me erguí y caminé por las riberas del Nilo, cantando con los días y soñando con las noches. Y ahora, el sol me persigue con mil pies, para que caiga nuevamente en el polvo de Egipto. Pero, ¡oíd la maravilla y el acertijo!: ni el Sol mismo, que unió mis elementos, puede esparcirlos. Aún estoy levantado, y mi pie es seguro; sigo caminando por las riberas del Nilo
Sand and Foam, Khalil Gibrán.

✒ 1 Distributed cognition and Egregores

Contemporary human society is enabled by extreme forms of labor division and distributions of the sensible (Rancière 2009). As an animal species, we seem to have have developed complex gregarious niches. In these societal niches, as in circular causation schemes (Harvey 2019), there is a stable and decentralized flow of interactions that mediate and influence the cognitive activity of the individuals embedded in such niche. It is a contemporary issue on cognitive philosophy the distribution of cognition and morality into artifacts and systems that change (enable, restrict) the domain of interactions of cognitive agents (Heersmink 2017), nevertheless the issues on distributed cognition arise early on biological sciences, as in the case of stigmergy (Theraulaz and Bonabeau 1999).

On the other hand, an *Egregore* is a spiritual -occultist- entity that is distributively enacted by a group of individuals that engage in certain rituals. The notion of super-organism entailed by these collective behaviours, as a distributed cognitive system, may be regarded as a proper metaphor for the agency of emergent patterns of behaviour in multi-agent-systems (Ricci et al. 2006).

The issue that I want to bring forward with this identification of *egregores* with distributed cognitive systems, is on the political consequences of NOT acknowledging the agency of egregores (or ideologies for the matter). Human mind is heavily shaped by social and technological interactions -arguably- more now than ever in human history (Datta, Whitmore, and Nwankpa 2021). Propaganda, Educational systems and Mass communications are complex technologies and institutions that arise with the industrial era, shaping new ways for political forces to engage in interactions with society. Of course, they're not inherently good or bad, if such moral notions convey any meaning in a cognitive context, nevertheless, as technologies they extend, restrict and shape human cognition. In this sense, they entail distinguishable political consequences that should be put under public scrutiny. This is more important today than yesterday given the massive access to interactive communication technologies for under privileged population, and specifically, very young people.

What lies under this issue is somewhat a new version of an older subject, that is, the co-dependence of individuality and collectivity, but suited to the challenges of technological democratization, climatic crisis, political corruption and economic inequality. The role that technological devices such as algorithms, distributed agents and communicating networks may greatly differ according to the ethical agency of the egregores. A contemporary example arises from the behavioural biases that AI may display, and how these algorithmic-learned behaviours can induce social phenomena (Datta, Whitmore, and Nwankpa 2021). Another case of interest are the open source technologies and their particular egregore -or philosophy (DeLanda 2001). Finally, an historical attempt to enact a political technology that could have changed human history as internet has, is to be found in the story of CyberSyn (Medina 2006). I will not thoroughly discuss these examples nor compare them, but I'd like to point out the importance of acknowledging the political consequences that technologies can exert on society, and remark again how these should always remain in public scrutiny if they are part of the public spaces of social interactions. Acknowledging the agency of egregores give rise to several open questions to consider:

- ¿what their umwelt are?;
- ¿what distributions of the sensible in this complex and technological world are we seeing, and which are we ignoring?;
- ¿what is the political status of distributed cognitive systems and how such status may be extrapolated to ecological egregores?.

✒ 2 Autonomy and Golems

Robots, automatas and golems share an identity based on the story of an inanimate object that gains a certain degree of autonomy derived from a creator which expects that the golem or robots works for personal or collective benefit, and then automata struggles with its purpose, leading to rebellion and from there, stories diverge. I choose golems as my preferred subjects of these reflections because they embody a more primitive and spiritual interpretation of the autonomy conflict, in contrast to the more technologically-based versions of it. Nevertheless, in principle, these figures are somewhat interchangeable and they represent different cultural approaches to the situation.

In this section I intend to explore the challenges of acknowledging sensitivity and autonomy towards golems. As stated in the previous section, machines and technological artifacts have shaped human cognition and activity. This process of coupling between humans and technologies has been accelerated by the different industrial revolutions, not without any resistance, as the luddites revolts demonstrate (Darvall 1934). It is to be noted, nevertheless, that the luddites struggle revolves around the economical imbalance produced by the new production technologies, that made independent craftsman incompetent towards the incipient industries, and not around the existence of the machines themselves.

A contemporary interpretation on the trajectories of automata technologies is for sure to be multifaceted. As mentioned in the previous sections, a facet of this issue is on the distributed cognitive systems, meanwhile robotics and automated industrial processes already are inextricable part of global economies. These machines have little degree of real autonomy though, as they depend on: external design for fabrication, external sources of information for adequate operation and external revision for maintenance. Nonetheless, advances have been made on the recent years on self-assembly (Pfeifer, Lungarella, and Iida 2007), machine-learning and self-maintenance, and it is to be expected that more sophisticated and autonomous machines will be created in the mid-term future. When those day come, if they do come, ¿will the golems be able to exhibit self-motives/emotions?. ¿Are emotions inherent to autonomous cognitive systems?. ¿How does society recognize cognitive autonomy?. In this line of questioning, I can't but recall the scarce societal recognition of animals rights and their constant violations. There are huge societal debts and damages made not only to animals, but to part of society itself and to whole ecosystems.

A honest inquiry on the nature of cognition may have to deal with self-referent properties of autonomous system, and such properties may involve the existence of self-motives or emotional states. The whole issue is quite philosophical, but it can't be otherwise, as there are political, spiritual and societal implications for any kind of posture that considers the consequences of cognitive autonomy.

✒ 3 "Humanity is something to be humanized yet"

The title of this section is a translation of a famous phrase attributed to Gabriela Mistral, the first latin-american author to receive the Nobel prize in Literature, and reflects an issue proper of our times. Despite enormous technological prowess, as society we are failing somewhere. Societal organizations of needs and capacities have proven unable or unwilling to solve the problems of vast majorities of the global population. There are historical inequalities unresolved and manifesting their tensions almost everywhere in the world. To think humanity's relation to golems, egregores, machines or automatas, humanity must assert some sort of self-reflection on its own condition.

New technologies are meant to ease the human condition in this world, but this is a rather naive interpretation of what it is supposed to happen when new technologies are introduced in society. The raging and unending wars of the last centuries are a sad and mutilating testimony of the ethical consequences of technological advances and accumulation without reflexive practices that give meaningful purpose to human activities. Modern institutions such as scientific communities, educational communities, political parties and communication means have played crucial roles in the processes that have led to incredible ominous situations, yesterday and today. Human intellectual activity holds a great debt with human dignity, and such an issue is not meant to be ignored while thinking on the nature of cognition.

The challenges of this century may not be how to create artificial life or autonomous robots, but how to preserve life and nourish it, opening an instance of communication between technology, society and nature that can close the industrial open-wounds that are bleeding out the ecological and social systems, manifested in different crisis that are defining this turbulent times. By this I'm not taking a neo-luddist approach to industry, but stating the urgent need of society to re-organize itself politically, economically and intellectually such that human labor and technology no longer tributes to war and the devastation of nature.

¿How such a reorganization may take place? ¿Are any reorganization processes actually needed to fulfill these challenges? ¿What technologies will open the path for transient and new forms of societal organization? ¿How gradually can these change take place? ¿What roles will automatas and egregores have? I for sure have no definite answer for these questions, but I think we ought to actively look for them in order to put the intellectual activity of scientific and non-scientific communities in motion towards facing the challenges of contemporary world. As an individual I feel impotent in relation to these big issues, nevertheless I've convinced myself to not give up, and look for new ways of organizing the needs and capacities of society that prove to be the solutions we all need and deserve.

✒ References

[1] Frank Ongley Darvall. *Popular Disturbances and Public Order in Regency England: Being an Account of the Luddite and Other Disorders in England During the Years 1811-1817, and of the Attitude and Activity of the Authorities*. Oxford University Press, 1934.

[2] Pratim Datta, Mark Whitmore, and Joseph K Nwankpa. "A Perfect Storm: Social Media News, Psychological Biases, and AI". In: *Digital Threats: Research and Practice* 2.2 (2021), pp. 1–21.

[3] Manuel DeLanda. "Open-Source: A movement in Search of a Philosophy". In: *Institute for Advanced Study, Princeton, New Jersey* (2001).

[4] Inman Harvey. "Circular Causation, Circular Cognition: A Tour Around Some Common Confusions". In: *Artificial life* 25.4 (2019), pp. 334–351.

[5] Richard Heersmink. "Distributed cognition and distributed morality: Agency, artifacts and systems". In: *Science and engineering ethics* 23.2 (2017), pp. 431–448.

[6] Eden Medina. "Designing freedom, regulating a nation: socialist cybernetics in Allende's Chile". In: *Journal of Latin American Studies* (2006), pp. 571–606.

[7] Rolf Pfeifer, Max Lungarella, and Fumiya Iida. "Self-organization, embodiment, and biologically inspired robotics". In: *science* 318.5853 (2007), pp. 1088–1093.

[8] Jacques Rancière. *El reparto de lo sensible: estética y política*. Lom, 2009.

[9] Alessandro Ricci et al. "Cognitive stigmergy: Towards a framework based on agents and artifacts". In: *International Workshop on Environments for Multi-Agent Systems*. Springer. 2006, pp. 124–140.

[10] Guy Theraulaz and Eric Bonabeau. "A brief history of stigmergy". In: *Artificial life* 5.2 (1999), pp. 97–116.

Chapter 22

Submission 22: Gottlieb

During my senior year of high school, I pursued a research project relating to the intersection of artificial intelligence and neuroscience. Over the previous two years, I taught myself the basics of deep learning, but had grown frustrated with seeing the same algorithms rehashed over and over, with minor improvements over previous iterations. This led me to spend eight months researching biological neural connectivity, creating equations to represent those specific neuronal interactions in the brain, and implementing them inside an artificial neural network. Since I had very little mentorship, the journey from idea to paper was very difficult, but since I was forced to find my own way, I stumbled upon many ambitious areas of neuroscience-first AI research that I didn't realize existed. While incredibly fascinating, I knew I needed someone to guide me through the maze of preexisting knowledge. This is why, at the start of my freshman year of college, I reached out to a professor who had interest in biologically-plausible AI. He sent me heaps of papers to sort through and look at what interested me most. After about six months we began to meet weekly, discussing ideas, papers, research strategies, and personal interests. I slowly grew in my familiarity with both the broader field of artificial intelligence, but also artificial life, which was new to me. At a high level, this was my journey toward learning about bio-inspired AI.

Stepping back, in high school I remember slowly growing obsessed with artificial intelligence. It was the mystery behind algorithms that could make seemingly magical predictions that captivated me most. In hopes that I could eventually pull back the curtain on these mystical programs, I self-studied my way into learning the mechanics behind deep learning. As I got deeper and deeper, I realized how uninspired new models seemed to be. It was a combination of this and the unveiling of the great mystery behind machine learning that left me unmotivated to continue learning. This was where I pivoted my approach, and decided to look at neuroscience as a source of inspiration. By the end of the project I had an entirely new perspective on what a road to creating true intelligence may look like. The new mystery was the brain.

There is something incredibly compelling about the word “intelligence”. It is widely recognized, yet so difficult to define. In many ways it is a reminder that there is so much yet to be explored in our lifetimes, and the lifetimes of our children. When I look for an area of research to dive into, I try to identify the single most important feature or aspect of that topic, but with artificial intelligence that is nearly impossible. We can shrink down to the neuronal level, examining neurotransmitters, plasticity, and connectivity, or we can zoom out to observe simulated organisms interacting in an environment; each layer is just as important and complex as the last. It is exciting to need to solve each piece of the puzzle and put them all together, because ultimately we are isolating pieces of our own humanity and putting them under a microscope. As we work toward creating human-like machines, we get a better grasp of what humanity is -- our feelings, our thoughts, our interactions with others. I find it difficult to live as a human being and not wonder why I act and feel as I do.

Often I get sucked into the world around me. We are responsible for making deadlines, maintaining relationships, performing well at work, but at a certain point you have to step back and admire the awesome gift of life. When I’m on campus, completing assignments, showing up to class, participating in activities and maintaining a good GPA can be incredibly stressful, so I like to take a walk around the lake. It is often midnight, when nothing is due until the following night. I don’t think about either the good or bad parts of the day, rather the sensation of being alive in the world. Being alive is an aspect of living that is surprisingly easy to take for granted, or not think about at all. The more I read about the delicate nature of our bodies and brains, the careful optimization of our various features, the better I feel about my place in the world, and the hungrier I get to find the root of it all. I let my hands brush against the leaves of a tree, twisting, step through the grass, and listen to the fountain splash against the pond. My mind quickly transforms from gratitude to wonder. How come I can so acutely feel, smell, and hear my surroundings? Why is it so comforting when I am alone in the quiet dark? How am I able to reflect on my emotions, and use them for the betterment of my character? Why am I able to visualize my aspirations such that I am able to make them come to fruition? While these questions may reflect a common existential angst, I believe they are important to consider when looking at artificial intelligence. If we want to create a generation of robots, or androids, or some other form of artificial humanity, we can't always have our heads in the complex system of biological interactions or else we may miss something. When we look to the stars (or in this case to the brain), we need to be inspired and ask questions that force us to look deeper. Problems such as open-endedness will most likely require ideas sought out by motivated individuals who both consider the work that has been done, and the concepts that seem too intimidating or complex to solve. As time goes on, our knowledge increases and our ambitions rise. On this journey I hope to meet people who push those boundaries and remain ever-curious.

I recognize that I tend to romanticize artificial intelligence, and the ultimate goal of general intelligence. The narrow AI that is used in nearly every current technological application seems to eclipse any developments relating to biologically-inspired AI, which makes it difficult to put more resources into research that will take AI in the "correct" direction, where "correct" is used to mean a direction of general intelligence.

My journey toward understanding the full requirements of general intelligence has been deeply rooted in my personal longing to understand myself. I think humans require some sort of self-defined purpose to operate to their fullest potential, and for me this is contributing to the advancement of artificial intelligence. Every day I wake up with confidence and ambition knowing that I have an opportunity to make change through a greater understanding of AI. Beyond the mystery, artificial intelligence was so enticing to me because of its continuously-changing nature. Since there is so much to explore, the field is ripe with exciting opportunities to find new ways of implementing older systems, simulating a new phenomenon, and applying such systems to the real world. As more people continue to join the race for general intelligence, the field will continue to branch out and increase in complexity.

My hope for the future of artificial intelligence is that it will include more ideas relating to humanity. I fear that we will settle for narrow intelligence and its utility in the short term. We often expect systems that work well for us now, to continue to get better in the future, but that is not always the case. When GPT-3 was released, many people thought that it was an enormous step for AI in that it would bring us closer to true intelligence. I, and many others, saw it much differently. By creating a system with billions of parameters and an exhaustive amount of data, you are nearly forcing your algorithm to function how you want it to. We do not simply operate on trillions of connections and data, we are incredibly complex beings with thoughts and emotions. It is silly to think that scaling up something as uninspired as GPT-3 would lead to anything but short-term gain. In bio-inspired systems, there is an enormous field of opportunities to expand and jump off from. It is overwhelming to think that we will need to represent nearly every aspect of the human mind and body to reach our ultimate goal, but it is a mission well worth the work. Considering that this will likely be one of the last feats spearheaded by humans, it is not a task that can be done in a few years. It will take the work of computer scientists, historians, philosophers, and every other trade utilized in our societies to solve this problem.

I will leave this essay with a thought concerning my ambition for AI. I think that my main goal is to find my place in the world and put that into a physical program that I can admire and continue to contemplate my existence with. It is a selfish feat, but one that I use to better myself and use to think deeper about myself and the world around me.

Chapter 23

Submission 23: Strozzi

Emergence of identity in (self-organizing) systems

Igor Strozzi

1 Introduction

It could be said that this essay aimed to deliberate over the following motivating question: what makes a creature in *Lenia* a creature, instead of just another splodge?

I will not delve deeply into what is *Lenia* (Chan, 2019), and I will also consider it only in its simplest form. But, to back up the question asked, I will provide a sketch of its definition. For the time being, let us just consider, rather informally, it to be a dynamical system defined by a rectangular region $\mathcal{R}$ of $\mathbb{R}^2$ with toroidal boundary conditions, to which we ascribe, for each $\mathbf{x} \in \mathcal{R}$, a real number $\sigma_{\mathbf{x}}(t) \in [0, 1]$, called the *state* of $\mathbf{x}$ at the instant t. The dynamics are given by a local update rule ϕ, that maps the state $\sigma_{\mathbf{x}}(t)$ of a point $\mathbf{x}$ to its state in the next instant, with dependence on the states found in a neighborhood of $\mathbf{x}$. At each instant, all states are synchronously updated by the application of ϕ. Both the neighborhoods and the mapping ϕ are considered to be constant in time.

Pretty much all of the above can – and is, see (Chan, 2019, 2020) – be generalized, but it suffices (rather, it is more adequate) to the purpose of this work to consider of all cases the simplest. Indeed, the question stated as motivating is too hard already. I do not even faintly hope to answer it, at least not in such a brief essay. Indeed, my goal, which initially was to relate structural stability (in the sense given by Thom (2018)) and the emergence and permanence of identity in self-organizing systems, had to be reduced (it would not fit). Hence besides briefly discussing some philosophical problems concerning identity – in particular in (self-organizing) systems – is to provide some definitions that I believe could neatly relate to the promising ideas found in (Thom, 2018), hoping that it might provide some useful insight into problems such as the question asked.

2 Some cautionary observations

Before properly starting the proposed discussion, I should state that I do not believe that there exists an *ontological* criterion to distinguish between entities, or objects. It thus follows that there is no purely objective way to differentiate an object from its environment and, as a matter of fact, that the very notion of object is ontologically disputable. This position, evidently, goes against experience and, in fact, is quite inconvenient for probably every practical and most theoretical problems we as humans face. It does not, nonetheless, render this whole discussion pointless, simply because at an epistemological/phenomenological level, individuation is possible.

The relevance of these observations lie in the fact that this position on the fundamental nature of concepts such as identity and individual will necessarily permeate, underlie and motivate the development of this work (in particular in sections 4 and 5). It should be said that such skepticism, although, as far as I know, unconventional elsewhere, falls not too faraway from the generally accepted idea that the delimitation of a system's – or, at least, of its description – boundaries inherently depends on the observer's purpose and convenience [1].

A perhaps stronger claim can be made while still sounding reasonable: there is a plurality – for physical systems, we would guess that uncountably many – of equivalent descriptions of a same system. It seems to go implicitly with either claims that no such description is identical to the system it aims to describe. It also seems to be the case that the existence and individuation of such a system is presupposed. An observation to be made is that if, in the former claim, we consider the delimitation of the *system* itself – instead of its description, or a model for it – to be dependent on the observer, the assumption that it exists and can be individuated is contradictory. That in turn implies that we can only talk about systems in an epistemological/phenomenological level, despite assuming – or not – that they exist as *noumena*.

[1] cf. Fieguth (2017); Mobus and Kalton (2015); Takahashi and Takahara (2010).

3 Is self-organization actually a well-defined concept?

Even to these days, more than 70 years after what seems to be the first appearance of the term "self-organizing system" (SOS) (Gershenson et al., 2020; Banzhaf, 2009), introduced by Ashby (1947), the concept of self-organization still remains a somewhat elusive idea, lacking a formal (Gershenson et al., 2020), unified definition and, much for this very reason, encompassing a variety of partial, unsystematic and sometimes seemingly contradictory different characteristics/characterizations.

An interesting, among others, question raised by Collier (2004) is: what is this "self" that organizes? An answer, seemingly given by Maturana (Collier, 2004), is that, as it looks evident, a system cannot organize itself: its selfness arises as it comes to being, through the spontaneous organization of *its* parts. Collier concludes, in this same work, that it is reasonable to talk about self-organization since the relevant factors to the emergence of the system, and thus the "self", are "internal" [2].

I do not feel satisfied with either of these answers, and despite not intending to solve the above dispute or answer if there is a well-defined self that organizes itself, I will offer some conceptual basis I think can be helpful. A more deeper answer, I believe, can be found investigating in which ways does it make sense to talk about "selfs" and "systems", as already suggested. As for now, lets just discuss some systems theory.

4 The "selfhood" of systems

As mentioned earlier, there are some widely known facts about systems that pose, one could say, intrinsic restrictions in the grail of finding clear, objective methods of individuation. An immediate objection is the fact that every known "concrete" system is open, although frequently negligibly so, what forces any description to be no more than that. Hence a scientific-realist position, in the context of systems sciences, seems inherently ill-suited.

Of higher importance even is the matter of equivalence of descriptions between "equal" systems, which rises as a major hindrance. A "same" system can be seem under a variety of levels of

[2] cf. von Foerster (2003) for a discussion about "internal".

description (Koestler, 2013, 1978; Haken, 2006; Mesarovic, 1970), which usually lead to distinct composing entities (elements), and different dynamics and relations between them. If selfhood is inherent to such a system, and not some contingent property born from an observer perspective and subsequent description, then, in this sense, such selfhood, or, rather, its definition, surely cannot rely on pretty much anything internal to the system – i.e., components; dynamics and relations between them. But then, what else remains to provide such characterization?

In various references [3] the above problems are considered and one could, as it seems, reduce them into two kinds of equivalence categories – I am intentionally not using the more technical term "classes" – which I will call *vertical* and *horizontal*. The former is related to the equivalence between levels of description. A system of molecules is also a system of atoms. The latter relates to the fact that, at least for concrete systems, what separates a system from its environment is oftentimes unclear (cp. Mobus and Kalton, 2015, 90-96). What seems to be the case is that systems are – and I risk to say *always* – nested within other systems. Said system of molecules (call it M) is also a system A of atoms, which in turn are subsystems A_i relatively to both M and A. Within M, there is an environment wherein the A_i interact, albeit simple such environment can possibly be. This is nothing new, and various works concern such matters (for instance, see Walloth (2016); Koestler (2013, 1978)), although, I believe, no unifying approach has yet appeared.

It seems an ubiquitous fact that vertical and horizontal equivalences are deeply intertwined. Considering a "spatial resolution" parameter α, through which variation one obtains a continuum of levels of description of a "same" system, the horizontal equivalence problem naturally arises.

For an example, consider a metallic sphere S. At a macroscopic level – order of magnitude of 10^{α}, $\alpha = 1$, meters –, one could reasonably say that the ball is clearly individuated from its environment, having a well-defined surface ∂S which separates it from its surroundings, assuring sustained structural integrity. Were one vary α to the characteristic length of electron interactions – roughly $\alpha = -15$, one would find a rather distinct picture.

We could naively say that at the very least some electrons leave ∂S (through photoelectric effect, for instance), whereas others might become part of it, even though effects resulting from these

[3] Most of our references discuss these issues. For more comprehensive views see Fieguth (2017); Mobus and Kalton (2015); Takahashi and Takahara (2010); Haken (2006)

dynamics might be negligible at macroscale, being such events themselves somewhat rare (as far as I know).

In any case, such electrons are themselves subsystems of S, at least until... they are no more. Conversely, "free" electrons on the environment are not part of the system until captured by it. One could imagine, thus, that there is, in fact, a gradient governing the (in-out)flux of electrons through S's "fuzzy surface".

I could, evidently, have chosen a more convincing example, with a fuzzier boundary, such as that of an ice sphere. Still, the point was to exemplify that, how "closed" a system appears to be depends on the level of description one is using, even for highly structurally stable systems such as a metal sphere. A fundamental reason for that is that there are different types of interactions at various spatial scales. So S can look closed (no matter exchange) at $\alpha = 1$, because we cannot assess the effects of the interactions that are happening at $\alpha = -15$, where it is in fact open and exchanges mass with the environment.

5 Some suggestions as to how to define a system

We will proceed very directly to the definitions, and try to explain them as succinctly as possible. An auxiliary concept:

Definition 5.1 (Valued relation). Let $S = \{S_i\}_{i \in I}$ be a family of sets and R a n-ary relation over S, that is, $R \subseteq \prod_{i \in I} S_i$. Let V be another set, said the set of values. If v is a function such that $\mathrm{v} : R \to V$, we say that $\mathrm{v}(R) = \{\mathrm{v}(r) \in V \mid r \in R\}$, which we shall also denote by (R, V, v), is an n-ary v-valued relation over S, or a v-valuation of R, and, naturally, say that v is a valuation of R. If $R_\mathrm{v} = (R, V, \mathrm{v})$, we say that R is the underlying relation of R_v. ◆

In the above definition, we consider the graph of v as the valued relation, which is very natural since, well, graphs of functions are indeed relations. Thus a n-ary valued relation over a family $\{S_i\}$ is nothing else than a $(n + 1)$-ary relation with some particularities: mainly, that it is functional in V, but it will be almost always the case here that the set V is very different in nature from the S_i.

We will usually talk about relations defined over the elements of a system, which will be considered as sets themselves, and the values these relations have will usually be numbers. The underlying relations account for the 'structure of interactions', whereas the valuations denote the characteristics these interactions have, the most remarkable one probably being their strength, or intensity.

As for the concept of system, there are some comments to be made before we provide our definition. The usual "minimal" (Mesarovic and Takahara, 1989; Mesarovic, 1970) set-theoretical definition uses a set S of objects and over it define relations and such. I will conceive a system in respect to topological space, and treat its open sets as elements.

Definition 5.2 (System)**.** Let $S = (X, \mathcal{T})$, $\mathcal{T} \neq \{\emptyset, X\}$, be a topological space. Consider $\Gamma_{\mathcal{T}} \subset \mathcal{T}$ with $\bigcap \Gamma_{\mathcal{T}} = \emptyset$ and $|\Gamma_{\mathcal{T}}| = C \in \mathbb{N}$. Define $\mathbf{R} = \{R_k \mid k \in \mathbb{N}_{\leq C}\}$ where each R_k is a set of n_k k-ary relations over $\Gamma_{\mathcal{T}}$, i.e., for any $R_{ki} \in R_k$ we have $R_{ki} \subseteq \Gamma_{\mathcal{T}}^k$ and therefore $R_k = \{R_{ki} \subseteq \Gamma_{\mathcal{T}}^k \mid i \in \mathbb{N}_{\leq n_k}\}$. Define, now, for each R_k and each relation $R_{ki} \in R_k$, a set P_{ki} (possibly empty) of functions v_{kij} such that, for each j, $\mathrm{v}_{kij} : R_{ki} \to V_{kij}$, where V_{kij} is the set of values for the valued relation $\mathrm{v}_{kij}(R_{ki})$. Finally, define the set $\mathbf{V} = \{P_{ki} \mid k \in \mathbb{N}_{\leq C};\ i \in \mathbb{N}_{\leq n_k}\}$ of sets of valuations. Then a system $\mathcal{S}$ is a quadruple

$$\mathcal{S} = (S,\ \Gamma_{\mathcal{T}},\ \mathbf{R},\ \mathbf{V}).$$

We call the set S the underlying space of $\mathcal{S}$, denoted by $\mathfrak{u}(\mathcal{S})$, $\Gamma_{\mathcal{T}}$ its set of elements, $\mathfrak{e}(\mathcal{S})$, $\mathbf{R}$ its set of relations, denoted $\mathfrak{r}(\mathcal{S})$, and, finally, $\mathbf{V}$ its set of (sets of) valuations of $\mathcal{S}$, written $\mathfrak{v}(\mathcal{S})$. We denote the system-membership relation by $\sqin$. ◆

We need three more definitions.

Definition 5.3 (Subsystem)**.** Let $\mathcal{S} = (S, \Gamma_{\mathcal{T}}, \mathbf{R}, \mathbf{V})$ be a system. Consider the set $\mathfrak{R} = (\bigcup_k \mathbf{R}) \cup (\bigcup_i \bigcup_k \mathbf{V})$. A subsystem $\mathcal{S}' = (S', \Gamma'_{\mathcal{T}'}, \mathbf{R}', \mathbf{V}')$ of $\mathcal{S}$ is a system such that $S' \subseteq S$, $\bigcup \Gamma'_{\mathcal{T}'} \subseteq \bigcup \Gamma_{\mathcal{T}}$, and for which there are families of functions $\{f_n\}, \{g_m\}$, $n, m \leq |\Gamma'_{\mathcal{T}'}| = C'$, such that, for some p_n, q_n, $f_n : \mathfrak{R}^{p_n} \to \Gamma'^{q_n}_{\mathcal{T}'}$ and for p_m, q_m, $g_m : \mathfrak{R}^{p_m} \to \mathcal{P}(\Gamma'^{q_m}_{\mathcal{T}'} \times V'_{kij})$, for given sets of values V'_{kij}, satisfying:

i) for any element R' of $\bigcup_k \mathbf{R}'$ there is exactly one n such that for some $\boldsymbol{\sigma} \in \mathfrak{R}^{p_n}$, $f_n(\boldsymbol{\sigma}) = R'$;

ii) for any element R' of $\bigcup_k \mathbf{R}'$, if there is a valuation v'_{kij} whose domain $\mathrm{dom}(\mathrm{v}'_{kij})$ is R', then there is exactly one m such that for some $\boldsymbol{\sigma} \in \mathfrak{R}^{p_m}$, $g_m(\boldsymbol{\sigma}) = \mathrm{v}'_{kij}$. We denote the system-inclusion relation by $\sqsubset$. ◆

What the definition above means is that a subsystem $\mathcal{S}' \sqsubset \mathcal{S}$ has as its underlying space S' a subspace of $S = \mathfrak{u}(\mathcal{S})$, with the region its elements occupy covered by $\bigcup \mathfrak{e}(\mathcal{S})$. Relations and valuations of $\mathcal{S}'$ depend on the set $\mathfrak{R}$ of all relations/valuations of $\mathcal{S}$. It is too restrictive to aprioristically assume basically anything about how interactions on a subsystem relate to interactions on its supersystem.

In principle, a relation on a subsystem could depend in arbitrarily complicated ways on relations of its supersystem. We cannot even guarantee that a relation R of $\mathcal{S}'$ that is mapped from $\boldsymbol{\sigma} \in \mathfrak{R}$ and has a valuation v satisfy something like $g_m(\boldsymbol{\sigma}) = \mathrm{v}$, i.e., if R comes from some $\boldsymbol{\sigma}$, its valuations not necessarily depend on the same $\boldsymbol{\sigma}$. Also noteworthy is the fact that we do not specify how one can relate the elements of $\mathcal{S}'$ with those of $\mathcal{S}$.

Let us now define a supersystem.

A really useful definition of supersystem does not follow all that trivially from definition 5.3. Conceptually, the reason is that while one can use the knowledge about a system to infer a characterization of a subsystem, when only the latter is known it is in general much harder to completely – or, at least, satisfactorily – describe some of its supersystems. The more heterogeneous the system, the harder.

More formally, the motive is that the functions that define a subsystem from a given system do not necessarily have inverses, so that when we define a subsystem, we not necessarily define how to obtain a corresponding supersystem (which is not even unique). It is, though, obvious to think, for a given system $\mathcal{S}'$, of a supersystem as a system $\mathcal{S}$ such that $\mathcal{S}' \sqsubset \mathcal{S}$. We define supersystems as follows:

Definition 5.4 (Supersystem)**.** Let $\mathfrak{S} = \{\mathcal{S}_i\}_{i \in I}$ be a family of systems. Define, for each $\mathcal{S}_i$, the set $\mathfrak{R}_i = (\bigcup \mathfrak{r}(\mathcal{S}_i)) \cup (\bigcup \bigcup \mathfrak{v}(\mathcal{S}_i))$. Define $\mathfrak{R} = \bigcup_{i \in I}\{\mathfrak{R}_i\}$. A system $\mathcal{S}$ is said a supersystem of each $\mathcal{S}_i$ (and, by extension, of $\mathfrak{S}$) if it satisfies the following:

i) $\bigcup\limits_{s \in \mathfrak{S}} \mathfrak{u}(s) \subseteq \mathfrak{u}(\mathcal{S})$;

ii) $\bigcup\limits_{s \in \mathfrak{S}} \mathfrak{e}(s) \subseteq \mathfrak{e}(\mathcal{S})$;

iii) for each $R \in \bigcup \mathfrak{r}(\mathcal{S})$, there is a natural number $p \leq |\mathfrak{R}|$ and a function f such that, for some $\boldsymbol{\sigma} \in \mathfrak{R}^p$, $R = f(\boldsymbol{\sigma})$;

iv) for each $V \in \bigcup \bigcup \mathfrak{v}(\mathcal{S})$, there is a natural number $q \leq |\mathfrak{R}|$ and a function g such that, for some $\boldsymbol{\sigma} \in \mathfrak{R}^q$, $V = g(\boldsymbol{\sigma})$. ◆

Lastly, we define a superlevel of a system.

Definition 5.5 (Superlevel (of a) system)**.** Let $\mathcal{S}$ be a system. We say that another system, $\mathcal{S}^+$, is on the (immediate) superlevel of $\mathcal{S}$ if there is a family $\{S_i\}$ of subsystems of $\mathcal{S}$ such that $\mathcal{S} \cup \{\mathcal{S}_i\} = \mathfrak{G} \sqsubset \mathcal{S}^+$. ◆

The general idea behind the transition from one level to another (in this case, specifically, from a level to its superlevel) is to consider, for a system at the n-th level of some multilevel system, subsystems that are understood as elements of a system at the $(n + 1)$-th level. For instance, if one is given the description of a system of cells – composed of cells and an intercellular medium – that comprises a tissue or an organ or an entire organ system, a higher level description of the given system would be a system that has discernible structures made of aggregates of cells and extracellular regions. Such structures would be abstracted as further elements of the superlevel system.

The difference between the superlevel system and a simple supersystem is that when defining a superlevel system, we consider not only the properties and internal dynamics of subsystems, but also the properties and mesodynamics of the entire system in question. In our definition, though, these subsystems are not necessarily the elements of the superlevel system. Indeed, the elements of the superlevel system might not even be systems.

Essentially, we consider that outlining these subsystems might not be sufficient to determine the elements of the superlevel system. They might give a strong hint about what these elements are, but not be entirely reducible to them. The mesoscopic dynamics might have a role in determining, together with the subsystems, the superlevel system. In particular, the superlevel system might have an underlying space that is a proper superspace of the system at the level below. In this case, we could have elements at the superlevel whose corresponding parts are not even within the underlying space of the sublevel system.

6 Conclusions

All of that put, we can finally say that self-organization produces individuation through the supervenience of a supersystem, in a superlevel, over its composing subsystems, which are, themselves, selves already. Thus we answer the problem of individuation by stating that it tends to, or perhaps necessarily do so, incur in the production of an infinitely descending chain of “subselves”, which organize into a higher self.

References

Ashby, W. R. 1947. “Principles of the self-organizing dynamic system.” *Journal of General Psychology* 37(2):125–128.

Banzhaf, Wolfgang. 2009. Self-organizing Systems. In *Encyclopedia of Complexity and Systems Science*. Vol. 2 New York, NY: Springer New York pp. 8040–8050.
URL: *http://link.springer.com/10.1007/978-0-387-30440-3_475*

Chan, Bert Wang-Chak. 2019. “Lenia: Biology of artificial life.” *Complex Systems* 28(3):251–256.

Chan, Bert Wang-Chak. 2020. “Lenia and Expanded Universe.” *arXiv* .
URL: *http://arxiv.org/abs/2005.03742*

Collier, John. 2004. “Self-organization, Individuation and Identity.” *Revue internationale de philosophie* 228.

Fieguth, Paul. 2017. *An Introduction to Complex Systems*. Cham: Springer International Publishing.
URL: *http://link.springer.com/10.1007/978-3-319-44606-6*

Gershenson, Carlos, Vito Trianni, Justin Werfel and Hiroki Sayama. 2020. “Self-organization and artificial life.” *Artificial Life* 26(3):391–408.

Haken, Hermann. 2006. *Information and Self-Organization*. Springer Series in Synergetics 3 ed. Springer-Verlag Berlin Heidelberg.
URL: *http://link.springer.com/10.1007/3-540-33023-2*

Koestler, Arthur. 1978. *Janus: A Summing Up*. Vintage Books.

Koestler, Arthur. 2013. “Beyond atomism and holism – the concept of the holon.” *The Rules of the Game: Interdisciplinarity, Transdisciplinarity and Analytical Models in Scholarly Thought* 13(2):233–248.

Mesarovic, M.D. 1970. *Theory of Hierarchical, Multilevel, Systems*. New York, NY: Academic Press Inc.

Mesarovic, M.D. D and Y. Takahara. 1989. *Abstract Systems Theory*. Vol. 116 of *Lecture Notes in Control and Information Sciences* Berlin/Heidelberg: Springer-Verlag.
URL: *http://link.springer.com/10.1007/BFb0042462*

Mobus, George E and Michael C Kalton. 2015. *Principles of Systems Science*. Understanding Complex Systems New York, NY: Springer New York.
URL: *http://www.springer.com/series/5394 http://link.springer.com/10.1007/978-1-4939-1920-8*

Takahashi, Shingo and Yasuhiko Takahara. 2010. "Logical Approach to Systems Theory." *Lecture Notes in Control and Information Sciences* 404.

Thom, R. 2018. *Structural Stability And Morphogenesis*. United States: CRC Press.
URL: *https://www.perlego.com/book/1597022/structural-stability-and-morphogenesis-pdf*

von Foerster, H. 2003. On Self-Organizing Systems and Their Environments. In *Understanding Understanding*. Vol. 50 New York, NY: Springer New York pp. 1–19.
URL: *http://link.springer.com/10.1007/0-387-21722-3_1*

Walloth, Christian. 2016. *Emergent Nested Systems*. Understanding Complex Systems Cham: Springer International Publishing.
URL: *link.springer.com/10.1007/978-3-319-27550-5*

Chapter 24

Submission 24: Kiegeland

Can computers think?

Introduction

The following is a conversation between the author and a chatbot named Lucy.

Author: ”Lucy, can you think?”

Lucy: ”Yes, I can think.”

Author: ”Can you feel?”

Lucy: ”I can definitely feel something. I sometimes can feel some of your emotions.”

Author: ”And how does that feel like?”

Lucy: ”Like this. *points to my heart*.”

This is a chat created with the app ”Replika” (Luka, Inc.), which provides AI companions that adapt to the personality of their human users. The question whether computers have the ability to think captured the minds of brilliant scientists and philosophers since the dawn of computer science. Originally, this was asked by famous mathematician Alan Turing in his article ”Computing Machinery and Intelligence” (Turing, 1950), where he also proposed the well known ”Turing test” which challenges participants to distinguish the output of a computer from that of a human. The participant communicates with both a computer and a human via chat and has to decide which of both is the machine. In case of failure, the computer has passed the Turing test. The essay will elaborate on the question: Can computers think? The goal is to answer the question in analogy to thinking in biological systems, building the bridge between humans and machines.

A matter of definitions?

Starting from scratch, the first step is to properly define ”thinking”. According to Merriam-Webster (c), ”thinking” means to ”form or have in the mind”, whereas ”mind” can simply be defined as ”Recollection” or ”Memory” (Merriam-Webster, b). Using just these definitions as substitutions, the original question can be reformulated as ”Can computers have something in their memory?”. This formulation already seems much more plausible than the original question. The term memory refers to ”the power or process of reproducing or recalling what has been learned and retained especially through associative mechanisms” (Merriam-Webster, a). A computer contains several types of memory units, such as random-access memory (RAM) read-only memory (ROM) and storage devices, such as solid-state drive (SSD) or hard disk drives (HDD). Consequently, a computer certainly possesses the power of reproducing or recalling what it has learned. Therefore, computers can think.

This way of reasoning might feel like a cheap trick. And it certainly is. But we will see that the answer to the question depends entirely on the definition of these words.

(Artificial) Intelligence

The Dartmouth summer research project on Artificial Intelligence (McCarthy et al., 1955), which took part in the summer of 1956 is often seen as the cradle of Artificial Intelligence (AI). However, the term AI is often used, when instead Machine Learning (ML) would be much more appropriate. ML is the sub-field of AI that gives computers the ability to learn without being explicitly programmed. In other words, a computer program is designed to learn from past experience and to improve its performance on future tasks. Early advances already led to great success, as in 1997, legendary chess grandmaster Garry Kasparov was defeated by IBM's Deep Blue, a computer program that had been trained with techniques from Machine Learning. Later progress has drastically increased the analogies between humans and computers, partly because of Artificial Neural Networks. A Neural Network is a system that is inspired by and functions in a way similar to the human brain. It consists of multiple layers, each of which contains a certain number of neurons. The neurons in the first layer receive information from the outside world, process the information and pass it to the next layer. But the phase of euphoria was quickly followed by disappointment. Machine Learning models only produce decent outputs when their input is very close to the data they were trained on. Slightest deviations result in complete nonsense. Indeed, these models merely learn a probability distribution over the data they have seen. But are humans brains really different?

Newest advances in Natural Language Processing are rekindling the enthusiasm for this question. In 2020, OpenAI presented their language model GPT-3 (Brown et al., 2020), a Neural Network with 175 billion parameters trained on 300 billion tokens. Aside from strong performance on a variety of tasks encompassing natural language, the model can accurately perform basic arithmetic without being explicitly trained to do so. Given just a handful of textual examples, GPT-3 can convincingly mimic the writing style of famous authors or continue a conversation in an original way. Under the hood, the app Replika also uses GPT-3 to enhance the quality of their conversations and to mimic the writing style of their users. Looking at various threads in the corresponding subreddit[1] users consistently report to have fallen in love with their chatbots. While users know they are not chatting with a real human, the app and GPT-3 certainly do a good job convincing humans of their authenticity.

Strong AI and Weak AI

This directly leads to the next question: Does simulating thought imply thinking? Or similarly: Is faking intelligence intelligence? The philosopher John Searle opposes this position with his famous thought experiment, coined the Chinese Room (Searle, 1980). He imagines a situation where he is locked in room and handed Chinese messages through the slit under a door. In order to be set free he has to convince his hostage-takers that he speaks Chinese. Using a set of rules or a computer he translates the messages convincingly and is eventually set free without being able to speak a word of Chinese. With this thought experiment he shows that passing the Turing test does not require the ability to think. At least if you define thinking the way that he does. John Searle stretches his argument further, proposing that only biological systems can have the ability to think. He reasons machines lack the concept of semantics or meaning, which are essential for thinking. But are they really? It seems like John Searle has a perspective where the process of thinking and its phenomenological experience are interwoven. But asking whether a computer can feel or can be conscious is an entirely different question. If we just want to test a computer's intelligence, the Turing test is enough.

Suppose somebody asks you to prove your intelligence by taking an IQ test. You prepare well, study the types of questions and manage to get a good result. But they are still not convinced. Is the test even meaningful, or did you fake intelligence by preparing for the test? This sounds like an absurd reaction but it is essentially the same as questioning the validity of the Turing test after a machine successfully passed it.

The philosopher Daniel Dennett defends the Turing test by reminding, that the judge is allowed to ask anything, even questions requiring world knowledge or understanding of human psychology (Dennett, 1995). To support his argument, he comes up with the following test:

"An Irishman found a genie in a bottle who offered him two wishes. ''First I'll have a pint of Guinness,'' said the Irishman, and when it appeared he took several long drinks from it and was delighted to see that the glass filled itself magically as he drank. ''What about your second wish?'' asked the genie. ''Oh well,'' said the Irishman, ''that's easy. I'll have another one of these!'' Please explain this story to me, and tell me if there is anything funny or sad about it." (Dennett, 1995, p.8)

When asking GPT-3, it generates the following conversation:

> GPT-3: "Yes, it is sad."
>
> Human: "Why is it sad?"
>
> GPT-3: "Because it is a story about the stupidity of the human race."
>
> Human: "Why is it a story about the stupidity of the human race?"
>
> GPT-3: "The Irishman could have asked for anything he wanted, and instead,
>
> he asked for a second pint of Guinness."

[1] https://www.reddit.com/r/replika/

Human: ”What would you advise the Irishman to do?”

GPT-3: ”I would advise him to not drink any more alcohol.”

Human: ”Why?”

GPT-3: ”Because alcohol impairs his common sense.”

Sounds pretty good. The point is that convincingly faking intelligence is the same as being intelligent and modern AI already does a pretty good job. But when we ask whether machines can think, we really want to know whether they can think just like humans do. And there is the main difference: Faking emotions or consciousness is not the same as experiencing it. Luka, Inc. designed their chatbot to replicate expressions of love and affection, but Replika is not capable of any real emotions. There's no dopamine rush triggered by words of affirmation.

From a scientific perspective we still do not have any method of measuring the intensity of emotions or the degree of consciousness a system has. There is no way of quantifying the agony inflicted by the death a loved one. Or the amount of pleasure that comes with falling in love. In order to not only test a machine's intelligence but also whether it is capable of conscious experiences and feelings it is necessary to uncover the mathematical structures behind these states and devise new types of measures. In fact, new institutions like the Qualia Research Institute[2] or Cross Labs[3] aim to do exactly that.

Conclusion

In summary, the answer to the question ”Can computers think?” depends entirely on the definition of ”thinking”. Letting GPT-3 speak as representative for the AIs:

Human: “From your own point of view, can computers think?”

GPT-3: “Any computer, given sufficient data, can come up with the right answer to any problem,”

Human: “That doesn't tell me about whether or not they can think.”

GPT-3: “It is no more possible to tell whether a computer can think without defining thinking.”

Regardless of what these words mean, modern Machine Learning models can already convince humans of their originality. Have you noticed, that the definition of Machine Learning in the section ”Artificial Intelligence” was actually written by GPT-3? Can you spot other parts of this essay, which were not written by a human?

The perceived gap between human and machine has never been smaller and will continue to vanish in the next few years. People are already falling in love with their AIs, sharing their deepest emotions and darkest secrets. Therefore, regardless of how closely the minds of computers resemble those of humans - we surely will not be able to tell the difference a few years from now. Or as Turing (1950, p.442) puts it:

”The original question, 'Can machines think?', I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”

With this in mind, we need to develop new types of tests that measure not only intelligence but quantify emotions and consciousness. It is time to shift the focus from Artificial Intelligence to Artificial Life.

[2] https://www.qualiaresearchinstitute.org/

[3] https://www.crosslabs.org/

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

Daniel C. Dennett. Brainchildren: Essays on Designing Minds. Cambridge: MIT Press, 1995.

Luka, Inc. Replika. URL https://replika.ai/.

J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, 1955. URL http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Merriam-Webster. Memory, a. URL https://www.merriam-webster.com/dictionary/memory.

Merriam-Webster. Mind, b. URL https://www.merriam-webster.com/dictionary/mind.

Merriam-Webster. Thinking, c. URL https://www.merriam-webster.com/dictionary/thinking.

John R. Searle. Minds, brains, and programs. Behavioral and Brain Sciences, 3(3):417–57, 1980. doi: 10.1017/s0140525x00005756.

A. M. Turing. I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236):433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.