A framework for designing compassionate and ethical artificial intelligence and artificial consciousness

Soumya Banerjee
University of Oxford
soumya.banerjee@maths.ox.ac.uk

Abstract: Intelligence and consciousness have fascinated humanity for a long time and we have long sought to replicate this in machines. In this work we show some design principles for a compassionate and conscious artificial intelligence. We present a computational framework for engineering intelligence, empathy and consciousness in machines. We hope that this framework will allow us to better understand consciousness and design machines that are conscious and empathetic. Our hope is that this will also shift the discussion from a fear of artificial intelligence towards designing machines that embed our cherished values in them. Consciousness, intelligence and empathy would be worthy design goals that can be engineered in machines.

Introduction

Consciousness has intrigued humanity for centuries. It is only now with the emergence of complex systems science and systems biology that we are beginning to get a deeper understanding of consciousness. Here we introduce a computational framework for designing artificial consciousness and artificial intelligence with empathy.

We hope this framework will allow us to better understand consciousness and design machines that are conscious and empathetic. We also hope our work will help shift the discussion from a fear of artificial intelligence towards designing machines that embed our values in them. Consciousness, intelligence and empathy would be worthy design goals that can be engineered in machines.

Architecture for artificial intelligence

In this section we outline a computational architecture for intelligent machines that can be considered to be conscious and empathetic. The key components of this architecture are described below and shown in Figure 1. Each component is also described in detail in the subsequent sections.

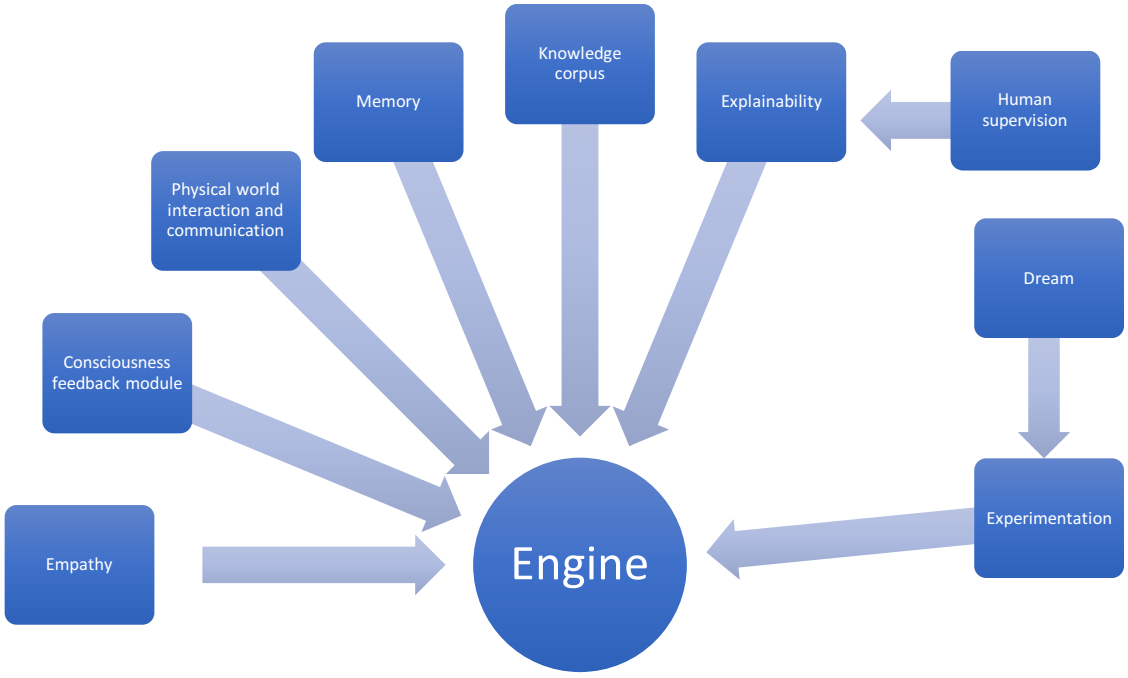

Figure 1. Computational architecture of artificial intelligence with consciousness and empathy.

1. Core computational engine

The core engine is a deep learning framework which is able to analyze data and formulate concepts. The different hidden layers of the neural network will represent concepts. For example, say the deep learning framework analyzes data from an oscillating pendulum. It will build an internal model representation of this data and discover an invariant like the period of oscillation (Figure 2) [1]. The deep learning framework will also analyze this data and form a concept of a period of a pendulum. This will also be reinforced by human input on what that concept actually means and by parsing the corresponding definition from an open-source knowledge corpus (like Wikipedia) (shown in Figure 2). We hypothesize that the more these machines connect these disparate sources of information and formulate concepts, the more knowledge it will have and the more intelligent and conscious it will become. These machines will also collaborate with other machines and human operators to communicate and share concepts; together they will build on these concepts to generate additional insight and knowledge. Our hope is that humans and networks of machines can collaborate, leading to a collective amplified intelligence.

2. Intelligence

The machine will assimilate knowledge from supervised human input and an open source knowledge corpus (like Wikipedia). It will use a deep learning framework to analyze data and formulate concepts (the different hidden layers will represent concepts). It will then

generate internal models, experiment with variations of these models, and formulate concepts to explain these (similar to a previous framework [1]).

Let us again assume that the machine is given data on an oscillating pendulum. It will form an internal model and mutate that model (using genetic programming or Bayesian techniques) whilst ensuring the model predictions match the empirical data (Figure 2). It will then analyze this to find invariants (quantities that do not change) like the period of oscillation of a pendulum. Performing this in a deep learning framework, the hidden layers of the network will represent the concept of a period of oscillation. These concepts will be fed into the explainability module which will translate this concept to a human understandable format. Human operators will then reinforce this concept with additional information and context (like more knowledge of physics or the concept of period in other dynamical systems). Operators can help generalize the concept of period of oscillation to other dynamical systems and help reinforce this concept (similar to how a teacher would teach this subject).

We note that this module can also be implemented in a Bayesian setting (Figure 2). The different models can be varied or mutated by using Markov Chain Monte Carlo in a non-parametric Bayesian model. Additional information can be provided from human operators using priors in a Bayesian model [2].

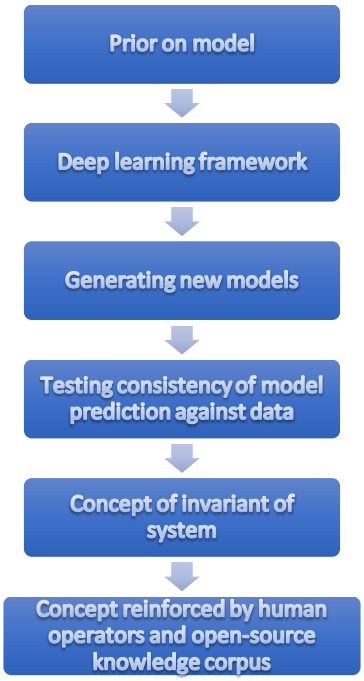

Figure 2. A Bayesian approach to artificial intelligence. The framework gets prior model specifications from human operators, mutates the models and checks consistency with empirical data. The hidden layers of the deep learning network represent a concept (for example, the invariant of a dynamical system). This concept is further reinforced by human operators (who help generalize that concept) and input from an open-source knowledge corpus.

3. Empathy module

The empathy module would build and simulate a minimal mental model of others (robots or humans) so that they can be understood and empathized with. We note that this is a computationally expensive module (see Section Designing Empathy in Machines). This would require enforcing some constraints on how much time is spent on processing that information.

4. Consciousness module

We define consciousness as what information processing feels like in a complex system (defined in detail in Section An Information Theoretic View of Artificial Consciousness). The consciousness module (Figure 4) is a deep learning network with feedback to analyze itself. The critical factor is that the module would feedback into itself. It would also need inputs from the intelligence module.

We argue, like others before [3], that consciousness is what information processing feels like. Due to learning and human feedback, consciousness can also be learnt over time (bicameral theory of mind and consciousness [4]). We hypothesize that this proposed engineered system would build up consciousness over time.

Communication with other robots and humans is also a critical component in order to build a collective intelligence. These machines will communicate with other artificially intelligent machines.

Finally, this module will have agents (like in agent based models) combined with machine learning or reinforcement learning. Consciousness will be an emergent property of this system.

5. Dream module

The dream module is a sandbox environment to simulate all other modules with no inhibitions. This is similar to DeepDream [5,6] with feedback into itself. This module also has connections to the experiment module.

6. Experiment module

The experiment module will play or experiment with systems in a protected sandbox environment (similar to another framework [1]). The input to this module would be data on a particular system of interest. The module would make a model, perturb this model, and observe how well is it consistent with data. The output of this module would be fed into a neural network to form concepts (the hidden layers of the deep learning network would represent concepts).

7. Explainability module

This module will run interpretable machine learning algorithms to allow human operators to understand actions taken by the machine and get insights into mechanisms behind the learning.

8. Unsupervised learning

We propose that these systems should also incorporate some information from curated repositories like the Wikipedia knowledge corpus, similar to what was done for the IBM Watson project.

9. Supervised learning

The concepts formed by the engine and the output of other modules will be tested by humans. Humans will interact with the explainability module to understand why these particular actions were taken or concepts formed. Human operators would ensure that these machines have a broad goal or purpose and that all their actions are consistent with some ethical structure (like Asimov's Laws of Robotics). Human operators will also try to minimize the harm to the machines themselves.

We expect that different machines will form distinct personalities based on their system state, data and human input. This is an opportunity for us to personalize these machines (if used in homes).

This step is also the most vulnerable; humans with malicious intent can embed undesirable values in machines and hence considerable care should be taken in supervising these machines.

10. Memory

Finally, the machine will have memory. It will record all current and previous states of its artificial neural network, learning representations and interactions with human operators. It will have the capability to play back these memories (in a protected sandbox environment), perform role-playing and simulate future moves. This will need to be done in the dream module. Our hope is that machines will be able to use past memories to learn from them, generate new future scenarios from past data and train on this data.

These "reveries" will allow the machines to effectively use data on past actions to generate new knowledge. This reverie architecture generates new knowledge by combining stored information with a capability to process that information (Figure 3). In a certain sense, these machines will be able to learn from past mistakes and adapt to different scenarios. We hypothesize that this will lead to higher levels of intelligence and ultimately lead to a form of consciousness.

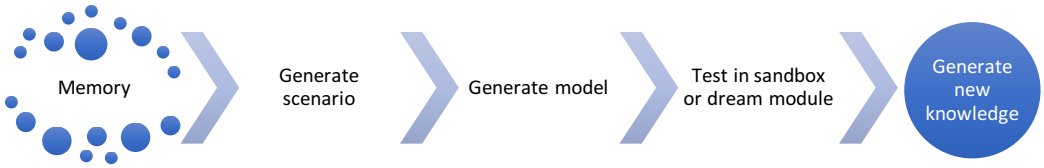

Figure 3. Using past memory to generate new knowledge. Past data is used to generate training data which is combined with new models. This yields new insights and knowledge into a system of interest. This "reverie" architecture generates new knowledge by combining stored information with a capability to process that information.

Consciousness, intelligence and life: perspectives from information processing

We hypothesize that consciousness, intelligence and life are different forms of information processing in complex systems [7]. Information and the computational substrate needed to process that information serve as the basis of life, intelligence and consciousness.

The minimal computational unit needed to create this artificial intelligence can be an artificial neuron, reaction diffusion computers [7,8] or neuromorphic computing systems [9]. Any of these computing substrates can be used to implement the architecture shown in Figure 1.

An information theoretic view of artificial consciousness

Consciousness is characterized by feedback loops in a complex system. Consciousness is what information processing feels like when there are feedback loops in a complex system that processes information [1, 3,10,11].

Consciousness also has been hypothesized to be an emergent property of a complex system [12]. It is like asking what makes water liquid; it is not only a property of the water molecule but also an emergent property of the entire system of billions of water molecules. Hence, we hypothesize that consciousness is also an emergent property of a complex information processing system with feedback.

Designing consciousness in machines

How can we encode these principles and design consciousness in a computer? A tentative basic definition of a conscious machine is a "A computing unit that can process information and has feedback into itself". An architecture of a consciousness module is presented below and shown in Figure 4.

Would a computer recognize that it is a computer? We can show a computer images of other computers to help it recognize itself (using deep learning based image recognition algorithms). We can also for example show the machine images of a smartphone, birds and buildings to reinforce the concept that it is not any of these things (non-self). Finally, we can design an algorithm to select out all images of non-self; all that remains is self.

This kind of an algorithm can be used to design a sense of self in machines. Such a supervised learning approach is similar to negative selection in biology [13] where the immune system learns to discriminate between all cells in the body (self), versus all that is foreign and potentially pathogenic (non-self).

A complementary approach is to exhaustively define all qualities that uniquely define self like the size of the computer, colour, amount of memory, identification marks, memory of past actions taken by this machine, etc. We can also show the computer an image of itself. All these attributes can then be coupled to a deep-learning based image recognition program or self-recognition program. The different hidden layers of a deep learning module would encapsulate the concept of self (based on images or other attributes).

Both these strategies can be combined to design a basic level of self-awareness and consciousness in computers (Figure 4).

We see some parallels to the turtle robots designed by William Grey Walter that could sense, move, and recharge [14]. These simple robots could follow a light source with a sensor, move around and then recharge its batteries when required. In one experiment, a light source was placed on top of the robot and the robot itself placed in front of a mirror. It was claimed that the robot began to twitch; William Grey Walter suggested that this robot showed a simple form of self-awareness [14].

We also argue that this is a basic level of consciousness and self-awareness. Similar principles can be used to design consciousness in machines.

Finally, we note that there are some well-known cognitive architectures that can be used to implement this form of artificial intelligence [15].

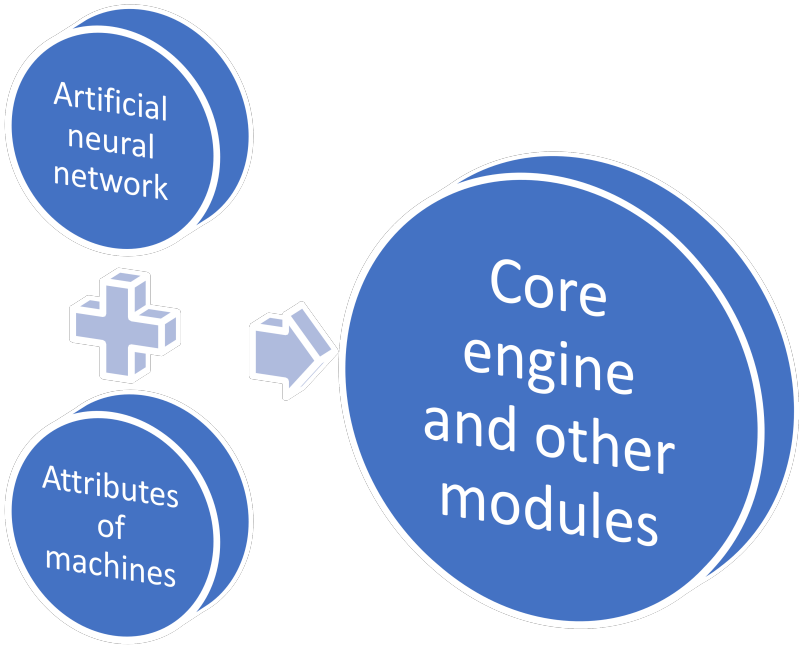

Figure 2. Architecture of the consciousness module. An artificial neural network analyzes the input and output of the core engine and all other modules (in Figure 1). This is combined with all the attributes of this particular machine (colour, physical characteristics, the fact that it is not a smartphone, it is not a bird, etc.). The hidden layers of the artificial neural network will represent the concept of self.

Human supervision and interpretability

Conscious machines will also need to explain themselves. The biological brain sometimes struggles to explain itself to others. Sometimes people find it hard to explain their actions or motives. Why do we love someone? Why do we feel afraid when we see a snake? The brain is a complex information processing system that does not lend itself very well to explanation.

Some progress has been made in making artificial neural networks interpretable [16,17]. These are approaches where artificial neural networks are turned on themselves to analyze their own actions. Similar techniques can be used to implement an explainability module (Figure 1) that can explain specific actions taken by the machine to human operators.

Making artificial intelligence interpretable or explainable will help human operators understand machines and help guide their training [18].

Designing empathy in machines

Empathy is when we try to deeply understand another person. The brain is like a Turing machine, and empathy is similar to running another Turing machine within it. Simulating a Turing machine with another Turing machine and asking the question whether it will ever halt is undecidable (Halting problem). We hypothesize that in general empathy is undecidable. It is also computationally expensive, which is perhaps why biological organisms do not have a lot of empathy.

Empathy is also intimately connected with a sense of self. Having a sense of self is essential for survival and maybe why evolutionarily it is important to have consciousness.
There are people called synesthete who have a heightened sense of compassion for other people. They feel intense emotions and empathy for other people to the point where human interactions exhaust them and they can become homebound. Essentially they are simulating other people and feeling what other people are feeling. They also find it difficult to separate their own self from other people.

Hence the reason we have a sense of self. We hypothesize that having a sense of self aids survival and delineates self from prey or predator. This may also be the reason we do not have a lot of empathy. If we did, we would not have a strong sense of self and may be at a selective disadvantage.

Empathy and consciousness are also related. The ability to run a simulation of what another person is feeling like (simulating another person's mental state) is empathy. Apart from being undecidable in general, empathy is also inversely related to a sense of self and hence maybe at a selective disadvantage.

Evolution may have decided that a lot of empathy is not good for individual survival. However, we have the unique opportunity of being able to engineer machines that have more empathy than biological organisms. We suggest the possibility of programming empathy in a computer. We may have to impose limits on how much time to simulate another person or another machine's state. In general this is undecidable, but we may be able to implement fast approximations.

Discussion

We present a computational framework for engineering intelligence, empathy and consciousness in machines. This architecture can be implemented on any substrate that is capable of computing and information processing [7,8,19,20-35].

We tentatively define consciousness as what information processing feels like in a complex system [3]. Consciousness is also like having a sense of self and is an emergent property of a complex information processing system with feedback.

Our proposed architecture for intelligent, conscious and empathetic machines will assimilate knowledge using supervised learning, form concepts (using a deep learning framework) and experiment in a sandbox. Our proposed machines will be capable of a form of consciousness by using feedback. They will also be capable of empathy by simulating the artificial neural state or conditions of other machines or humans.

We hypothesize that empathy and emotions have been pre-programmed over evolution. Empathy may confer an evolutionary advantage. We also recognize why too much empathy can be a disadvantage. Empathy can be achieved in robots through operator training (reinforcement learning) and allowing machines to analyze the artificial neural state of other machines or personal history of humans. We have the unique opportunity of being able to engineer machines that may have more empathy than biological organisms.

Communication technologies and human supervision can help accelerate the onset of artificial consciousness as was hypothesized to have led to the emergence of consciousness in humans [4]. Consciousness can be learnt over time as has been hypothesized before [4]. We suggest a computational approach to engineer this in machines with close human supervision.

Ultimately, we may be able to engineer higher levels of consciousness. More levels of feedback and more complexity in information processing may lead to higher levels of consciousness. The union of machine intelligence with our biological intelligence may also give us access to higher levels of consciousness.
Computing paradigms that are not constrained by physical space or have different computing substrates (as proposed in different information processing systems [7,36] and in biology [20-35]) may be capable of higher levels of consciousness. Our greatest contribution as a species may be that we introduce non-biological consciousness into the Universe.

We foresee a number of dangers. The scope of this computational framework (presented in Figure 1) is very broad and maybe currently be beyond the reach of individuals and only be feasible by giant corporations. Malicious corporations and conglomerates of individuals may misuse such an artificial intelligence, by for example failing to invest in empathy. It may be worthwhile to create non-profits that advocate for designing empathy in future intelligent machines and also educate the public about the potential benefits of such technologies.

Another danger is that we mistreat these artificial creations. What ethics might we need to create for conscious machines [37]? Would it be ethical to turn off or destroy such an artificially

intelligent and conscious being? The creation of artificially conscious and intelligent machines will challenge us to come up with new ethical structures. It may be the first time that consciousness would have been engineered rather than self-emerge and these beings would deserve as much sympathy as we show towards other species.

We hope that this framework will allow us to better understand consciousness and design machines that are conscious and empathetic. We hope this will also shift the discussion from a fear of artificial intelligence towards designing machines that embed our cherished values in them. Consciousness, intelligence and empathy would be worthy design goals that can be engineered in machines.

Acknowledgements
The author would like to thank Irene Egli, Chayan Chakrabarti, Tarakeswar Banerjee, Kalyani Banerjee, Linda Iglehart, Peter Grindrod, Franziska Diethelm and Joyeeta Ghose for fruitful discussions.

References

[1] Lipson, H. and Schmidt, M. Distilling free-form laws from experimental data. Science 234 (Apr. 3, 2009), 81–85.

[2] Soumya Banerjee, Melanie Moses, Alan Perelson (2017). Modelling the effects of phylogeny and body size on within-host pathogen replication and immune response, 14:20170479. http://doi.org/10.1098/rsif.2017.0479

[3] Tegmark, Max. Consciousness as a State of Matter. Chaos, Solitons & Fractals 76 (2015): 238-270

[4] Jaynes, Julian. The origin of consciousness in the breakdown of the bicameral mind. Houghton Mifflin Harcourt, 1976.

[5] https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html, URL accessed 30th December, 2017

[6] Banerjee, Soumya. Beauty of Life in Dynamical Systems. 2017, https://www.researchgate.net/publication/320011406_Beauty_of_Life_in_Dynamical_Systems

[7] Soumya Banerjee, A Roadmap for a Computational Theory of the Value of Information in Origin of Life Questions, Interdisciplinary Description of Complex Systems, 14(3), 314-321, 2016

[8] Adamatzky, Andrew, and Benjamin De Lacy Costello. Experimental logical gates in a reaction-diffusion medium: The XOR gate and beyond. Physical Review E 66, no. 4 (2002): 046112.

[9] Schemmel, Johannes, Daniel Briiderle, Andreas Griibl, Matthias Hock, Karlheinz Meier, and Sebastian Millner. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on, pp. 1947-1950. IEEE, 2010.

[10] Kaku, Michio. The future of the mind: The scientific quest to understand, enhance, and empower the mind. Anchor Books, 2015.

[11] Chalmers, David J. The character of consciousness. Oxford University Press, 2010.

[12] Godel, Escher, Bach: An eternal golden braid. Douglas R. Hofstadter. (Originally published by Basic Books, 1979.) Basic Books (1999).

[13] Janeway, C., Travers, P., Walport, M., & Shlomchik, M. (2005). Immunobiology: The immune system in health and disease. New York: Garland Science.

[14] http://rutherfordjournal.org/article020101.html, URL accessed December 30th 2017

[15] https://github.com/SoarGroup, URL accessed December 30th 2017

[16] Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. Restricted Boltzmann machines for collaborative filtering. In Proceedings of the 24th international conference on Machine learning, pp. 791-798. ACM, 2007.

[17] Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015).

[18] http://nautil.us/issue/40/Learning/is-artificial-intelligence-permanently-inscrutable, URL accessed December 30th 2017

[19] Vallverdu, Jordi, Oscar Castro, Richard Mayne, Max Talanov, Michael Levin, Frantisek Baluska, Yukio Gunji, Audrey Dussutour, Hector Zenil, and Andrew Adamatzky. Slime mould: the fundamental mechanisms of cognition. arXiv preprint arXiv:1712.00414 (2017).

[20] L Drew, F Stephanie, B Soumya, C Candice, C Judy, M Melanie. A spatial model of the efficiency of T cell search in the influenza-infected lung, Journal of Theoretical Biology 398 (7), 52-63, 2016

[21] Banerjee, S., Guedj, J., Ribeiro, R. M., Moses, M., & Perelson, A. S. 2016. Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-.http://doi.org/10.1098/rsif.2016.0130

[22] Science and technology consortia in US biomedical research: A paradigm shift in response to unsustainable academic growth. Curt Balch, Hugo Arias-Pulido, Soumya Banerjee, Alex K. Lancaster. BioEssays 37 (2), 119-122, 2015

[23] Soumya Banerjee and Joshua Hecker. A Multi-Agent System Approach to Load-Balancing and Resource Allocation for Distributed Computing, arXiv preprint arXiv:1509.06420, 2015

[24] Soumya Banerjee and Melanie Moses. Immune System Inspired Strategies for Distributed Systems. arXiv preprint arXiv:1008.2799, 2010

[25] Soumya Banerjee and Melanie Moses. Scale Invariance of Immune System Response Rates and Times: Perspectives on Immune System Architecture and Implications for Artificial Immune Systems. Swarm Intelligence 4, 301–318 (2010). URL http://www.springerlink.com/content/w67714j72448633l/

[26] Soumya Banerjee, Jeremie Guedj, Ruy Ribeiro, Melanie Moses, Alan Perelson (2016). Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-. http://doi.org/10.1098/rsif.2016.0130

[27] Soumya Banerjee. 2009. An Immune System Inspired Approach to Automated Program Verification, arXiv preprint arXiv:0905.2649, 2009

[28] Soumya Banerjee. Scaling in the immune system, PhD Thesis, University of New Mexico (2013)

[29] Soumya Banerjee. A Biologically Inspired Model of Distributed Online Communication Supporting Efficient Search and Diffusion of Innovation. Interdisciplinary Description of Complex Systems 14 (1), 10-22, 2016

[30] Soumya Banerjee. A computational technique to estimate within-host productively infected cell lifetimes in emerging viral infections. PeerJ Preprints 4 (e2621v2) 2017

[31] Soumya Banerjee. An artificial immune system approach to automated program verification: Towards a theory of undecidability in biological computing. PeerJ Preprints 5 (e2690v1) 2017

[32] Soumya Banerjee. An artificial immune system approach to automated program verification: Towards a theory of undecidability in biological computing. PeerJ Preprints 5 (e2690v1) 2017

[33] Soumya Banerjee. An Immune System Inspired Theory for Crime and Violence in Cities. Interdisciplinary Description of Complex Systems, 15(2):133-143, 2017

[34] Soumya Banerjee. Analysis of a Planetary Scale Scientific Collaboration Dataset Reveals Novel Patterns. arXiv preprint arXiv:1509.07313, 2015

[35] Soumya Banerjee. Optimal strategies for virus propagation. arXiv preprint arXiv:1512.00844, 2015

[36] Clarke, Arthur Charles. The wind from the sun. Vista, 1996.

[37] Sandberg, Anders. Ethics of brain emulations. Journal of Experimental & Theoretical Artificial Intelligence 26, no. 3 (2014): 439-457.

[38] https://www.ted.com/talks/david_chalmers_how_do_you_explain_consciousness, URL accessed December 30th 2017

[39] Minsky, Marvin. The emotion machine. New York: Pantheon 56 (2006).

[40] What is it like to be a bat? Thomas Nagel, The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450

[41] M. Oizumi, L. Albantakis, G. Tononi, From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0, PLoS computational biology 10 (2014) e1003588.