A path to Al

Ion Dronic dronic@narrafy.io www.narrafy.io

Abstract

To build a safe system that would replicate and perhaps transcend human-level intelligence, a system composed of three basic modules: objective, agent, and perception is proposed for development. An objective module would make sure the system acts in humanity's interest, not against them and would have two components: a network of machine learning agents and a cryptographic ledger. The network of agents would benefit from a form of semiotics, known as deconstruction to address the problem of value alignment [4]. A decentralized cryptographic ledger to propose a mechanism to mitigate the existential threat such a system would pose to humanity. To build the agent module that would act intelligently on the world, the concept of Dyna [11] would benefit from a treatise in sociology to build up the theoretical framework needed for the missing link of artificial general intelligence - the world simulator. A perception module that would estimate the state of the world, would benefit from existing deep learning algorithms enhanced by a new paradigm in hardware design - a quantum computer. This paper describes a way of how such a system could be built, analyzing the current state of the art and providing rather alternative directions for research than concrete, industry-ready solutions.

1 Introduction

The fig tree is pollinated only by the insect Blastophaga grossorum. The larva of the insect lives in the ovary of the fig tree, where it gets its food. The tree and the insect are thus heavily interdependent: the tree cannot reproduce without the insect; the insect cannot eat without the tree; together, they constitute not only a viable but a productive and thriving partnership as well. Evolution proves that a cooperative "living together in intimate association, or even close union of two dissimilar organisms" is possible. It is called symbiosis and it was J. C. R. Licklider¹ who proposed a man-machine symbiosis back in 1960, a thriving partnership between an artificial intelligence and humanity [1]. The latest progress in the artificial intelligence development would foster a hope that such a system could be built within this generation's lifetime. Al community largely agrees that an intelligent artificial system would emulate the human brain and

would be composed of 3 basic modules [2]:

- **objective module** would pick an objective and estimate if the system is satisfied or not.
- **agent module** would generate actions that are going to act on the world: prediction, planning, reasoning, memory etc.
- **perception module** would estimate the state of the world: video, audio, speech etc.

A thought experiment would explain how such a system would work.

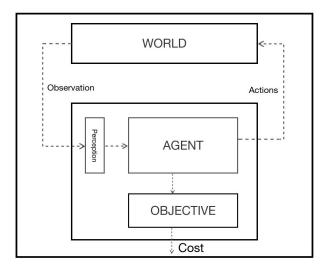


Fig.1

For example: Let's assume the system's **objective** is to find a job. The **agent** would start by writing a CV; search for available positions; send out emails. At the same time his **perception** module would **observe** the **world** for feedback: is there anyone replying to its emails. If someone replied, the agent would seek to schedule an interview, pass it and get the job. It would repeat this sequence of **actions** until it would eventually achieve the objective with a minimal **cost** - time and effort spent.

¹ J. C. R. Licklider - was an American psychologist and computer scientist who is considered one of the most important figures in computer science and general computing history. He is the first director of ARPA, known as DARPA - the military program that financed the internet.

Over the past few years, due to Deep Learning [3], important progress was made in the areas of speech recognition, computer vision, machine translation, reasoning, reinforcement learning (playing games), robotics and control. One can notice that most of the progress was made towards building a module that would estimate the world's current state — the perception module. The agent and objective modules are also in development but with different degrees of progress. The order of the implementation is also very important. If the agent module is developed first and the system's objective is misaligned, humans are in great danger (Bostrom 2014). An artificial intelligence that would achieve a human level of intelligence would quickly transcend it. And it is very common for humans to destroy other species ecosystems and build their own — only because they are the most intelligent species. There are currently no indications that an AI system would behave the same, but it's better to be safe than sorry.

2 Objective Module

To mitigate any of the doomsday scenarios an objective module composed of two submodules: value and control should be developed first. The module would have to answer questions such as how to make sure the system would uphold the same values as people; how to govern an intelligence that would be smarter than all human intelligence put together; how to specify an ultimate goal the system would need to achieve with a minimal cost etc.

2.1 Value Module

The value module or the "heart" of the system would make sure an AGI would uphold the same values as people. There are several ideas under development that address the "value alignment" problem. The approach called "Cooperative Inverse Reinforcement Learning" [4] is the one that got traction in AI community. It is the idea that the best source of information about what people value is human behavior. A slightly modified implementation of the concept was developed by two leading research groups in the industry, OpenAI, and DeepMind [5]. They proposed an algorithm which can infer what humans want by being told which of two proposed behaviors is better. The algorithm would remove the need for a person to write complex goals and learn what a person wants by observing what it prefers. In other words, it would learn what people value by observing their behavior. While the implementation of the algorithm is brilliant, the theoretical background might raise some concerns. What happens if a person picks the "lesser of two evils" and an agent learns from that choice. A choice that does not reflect a person's values, but seemed the best option at a particular point in time. Hence more, a person's day to day behavior is

constrained by the norms of the society he is living in, something acceptable in the western hemisphere of the planet cannot be tolerated in the Muslim world or far east. Even in fairly similar socio-cultural aggregations behavior is rarely a source of human values. A typical family therapy case would perhaps better illustrate the concern.

It is often that parents who have problems in dealing with their children come into counseling. They would tell stories about how they can't help themselves but shout at their children all the time. They may have enacted many moments of love and care. Yet if the story of themselves as bad parents is sufficiently strong, then these moments of love and care may be 'written out' - no significance is attributed to them and they wouldn't talk or show them in their behavior. [7]

If one would allow an intelligent system to learn what people value by observing their behavior, it might happen that the agent would learn that shouting at children is something acceptable and there is a value encoded in this behavior. Of course, this is an extreme case but human behavior cannot be a relevant source of what people value (Sachs 2017)[6]

It is a challenge to learn what people value and a great opportunity. Humanity has a chance to reinvent moral thinking and incorporate the knowledge that is often ignored or forgotten. To address the problem of "value alignment" an approach that would look beyond human behavior is proposed for development.

2.2 Narrafy

From Derrida ("Of Grammatology" 1967) it is not possible to talk about anything without drawing out what it is not. Every expression of life is in relation to something else. Expressions don't have an intrinsic relationship with the thing described. Words help 'make things out' by distinguishing one thing in relation to another. Words are relational (unlike 2D or 3D things which are more representational) and are always based on the distinction with that which it is not. (e.g. "injustice" only has significance in relation to "justice", distinguishing "despair" depends on an appreciation of "hope", "darkness is the mere absence of light").

The process is known as deconstruction (Derrida 1967), a form of semiotic analysis would guide the research towards finding a solution for the difficult tasks of learning what people value. The idea is simple - learn people's day-to-day problems to figure out what are the values hidden behind the problems. The invisible side of problems is what Michael White [8] calls 'the absent but implicit'. Deconstruction principles could be implemented into today's therapeutic mental health care via an artificial narrow intelligence - a system trained to perform a task in one specific domain. Industry examples would include

Deepmind's AlphaGo, an agent that beat a human at the ancient Chinese game of Go; IBM Watson an agent that won Jeopardy, a quiz based TV show. Narrafy is the proposal for a machine learning agent trained in counseling and therapy. The system would be exposed to the outside world through a chatbot interface.

2.3 Control Module

The control module is the second module of the system's objective. It would implement a mechanism to govern and control an artificial intelligence system. It is a safety measure to make sure the system won't pose an existential threat to humanity. Evolution's implementation of such a mechanism is the vascular system. A stream of cells (known as blood) delivers oxygen to the human brain. The brain needs oxygen for metabolism. The moment metabolism stops, the brain ceases to operate. It is the vascular system that controls the system's brain. It will be a stream of cryptographic tokens to play a similar role. It would carry keys to decrypt data needed for the intelligent system to run its own metabolism what that can be is further discussed, but in this way the system would need people to produce data - the oxygen of the system; people would need the intelligence to thrive - the brain of the system; if people would stop producing data the brain would cease to exist, but the opposite is not true. It is not a particularly new idea in the AI community. Carl Schumann² of Oxford University proposed a similar concept: "One could build an AI that places final value on receiving a stream of "cryptographic reward tokens." These would be sequences of numbers serving as keys to ciphers that would have been generated before the AI was created and that would have been built into its motivation system." [9] Human body uses bone marrow to produce **blood cells**, a cryptographic ledger could be used to produce tokens and achieve a similar function.

2.4 Token

Cryptography ledgers use tokens to anonymously prove ownership of an asset. It could be money, data or something else. A token could be used to prove ownership of real world-data a person would create during his lifetime and could be fed into the intelligent system. This data would reflect an honest, unfiltered perception of a person's reality, description of his world, people and problems a person faces on a day-to-day basis or a person's **snapshot of reality**. The token would have a dual purpose in the system - a mechanism to control the intelligent system and a mean to reward people for renting their personal data. It is important to acknowledge that the ownership of data is shared among the actors that participated in its creation: people and machine learning agents. Together,

² Carl Shulman is a Research Associate at the Future of Humanity Institute, Oxford Martin School, Oxford University

they would establish a symbiotic partnership and create the stream of cryptographic tokens or the "blood" of the system.

2.5 Ledger

A distributed cryptographic ledger could be used to store and encrypt information. The intelligent system would be conditioned to use data encrypted in the ledger and would need to pick one of the following scenarios to access it: rent, seize or bypass.

- **rent**. The best way to access data would be to rent them from people. The agent would have to figure out some sort of economic activity to be able to pay for the decryption keys.
- seize. Another way to get access is via an attack. It can try to seize the keys from people's wallets. To prevent the attack scenario, all network nodes participants would hold their keys in their secure wallets online, offline or on paper. As the Al development progresses the network nodes would be encouraged to hold the keys mostly offline and a certain percentage of nodes would be required to have their keys on paper storage.
- bypass. The agent can try to employ powerful computers to bypass data encryption. Given the fact the agent would transcend human intelligence, there is no doubt that it would find a way to do it. There are no ciphers that can't be decrypted, the only thing one can do is to make the decryption process hard and cost ineffective. And this is where the decentralized nature of the ledger would leverage its full potential. The intelligent agent would need to decrypt millions of nodes simultaneously, which would require high bandwidth and a large amount of energy. The community would notice the attempt and launch countermeasures. An example of such a measure would be something called a "network fork" [10] a way to roll back to a previously stored checkpoint and continue on a separate chain with the safe version of the system. The agent would then lose access to the ledger and the ability to run computations. The agent would always compete with a younger version of itself. It would be in his best interests to cooperate and not launch such attacks.

3 Agent Module

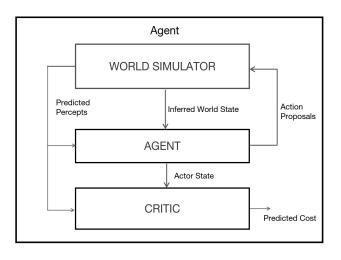
The agent module would be the module to act intelligently on the world. To understand why such an agent wasn't built yet, it would make sense to look at the obstacles current research groups are facing. There are many things that are missing, but perhaps the most important one is how to teach a machine common sense. And what is common sense in general? Perhaps the best analogy for

common sense is the human's brain ability to fill in the blanks. It can fill in the retina's blind spot; fill in missing segments in the text, missing words in speech; infer the state of the world from partial information; infer future from the past and present; infer past events from the present state; predict the consequences of actions leading to a result etc. Human brains are in fact prediction engines and one can say that **prediction** is the essence of intelligence (LeCun 2017)[2].

3.1 Dyna

Dyna is an integrated architecture for learning, planning and reacting [11]. It is the old common sense idea that predicting is trying things in your head using an internal state of the world. For an agent to act intelligently it would need an internal world simulator, an actor and a critic function. For example:

A super-intelligent agent that looks for a job would seek the company it would want to work for. Then, it would run different scenarios with action proposals according to its internal state of the world to predict what product the company is hiring for. When such a product is figured out it would run another sequence of actions to develop it. When the product is developed, it would



contact the company and negotiate rather an acquisition than a job interview. It would maximize the output and minimize the cost.

The module would need to have a predefined sequence of actions to run; measure the expected error using a critic function; adjust the action proposals to optimize the critic and train itself to produce better and better outcomes. The industry developed ways to build most of the agent components with one exception - the world simulator, which is believed to be the missing link of an artificial general intelligence (LeCun 2017). The concept of Dyna would be further developed to propose a solution to the world simulator problem.

3.2 World Simulator

The progress of the current society is largely attributed to several revolutions that happen during our history. The first and by far the most important is the cognitive revolution [17]. While its implications were profound and go way beyond the

scope of this paper, one particular development might contribute to a better understanding of how our world works. The ability to create fictions, imagine things that don't really exist, is believed to have enabled millions of people to cooperate towards achieving common goals. Things like nation-states, law systems, limited liability companies that weren't possible before, suddenly became possible. Even though these abstract concepts can be far from objective reality, they produce real-world repercussions. For example, the nation-state is a particularly new concept, there was no nation states few thousands of years ago, but it produces real 'realities' like customs, passports, citizenships, even wars sometimes. Over time people developed an incredibly complex network of these types of fictions. In academic circles, they are known as 'social constructs' or imagined realities [12]. To build a world simulator, there is a need to also build this extra layer humanity developed on the top of existing, observable, 'real-world'. A thought experiment [13] would explain better how this layer was constructed.

Imagine two survivors of some ecological disaster coming together to start a new society. Imagine that they are a man and a woman who come from very different cultures. Even though they have very little in common they would need to coordinate their activities in order to survive. As they do this, some agreed-upon habits and distinctions will emerge: certain substances will be treated as food, certain places found or erected to serve as shelters, each will begin to assume certain routine daily tasks, and they will almost certainly develop a shared language. They will always be able to remember, "This is how we decided to do this". They will carry some awareness that other possibilities exist. However, even in their generation, institutions as "childcare", "farming" and "building" will have begun to emerge. For the children of the founding generation, "This is how we decided..." will be more like "This is how our elders do it", and by the third generation it will be "This is how it's done". Mothers and farmers and builders will be treated as always-having-existed types of people. The rough-and-ready procedures for building houses and planting crops that our original two survivors pieced together will be more-or-less codified as the rules for how to build a house or plant corn. By the fourth generation of this imaginary society, "This is how it is done" will have become "This is the way the world is, this is reality".

As Berger and Luckmann put it "An institutional world...is experienced as an objective reality." They propose several processes that are important in the way any social group constructs and maintains its knowledge concerning "reality": typification, institutionalization, and legitimation.

- typification is the process through which people sort their perceptions into types or classes of objects.
- **institutionalization** is the process through which institutions arise around sets of typifications: the institution of motherhood, the institution of law, etc.

Institutionalization helps societies maintain and disseminate hard-won knowledge.

- **legitimation** is the word for the processes that give legitimacy to the institutions and typifications of a particular society.

Reification is the term to encompass the overall process of which the other three are parts. One could see the reification process as a nested hierarchy of social construction concepts. Accidentally or not, deep learning a field of machine learning that over past years gave impressive results in teaching machines recognize things it didn't see before (a "typification" process), would also represent the world as a nested hierarchy of concepts [3]. And since both the objective world and the extra layer people build on top of it, seem to be represented as a hierarchy of concepts, one can draw a conclusion that reality as people experience it could be reproduced using deep learning algorithms. The process of typification, institutionalization, and legitimation could be represented as levels of this very deep neural net, one calls reality. And, if this is the case and there is a theoretical framework, a world simulator that would mimic a reality would mainly depend on enough computational power and further research in this direction.

3.3 Reality Deconstruction Algorithm

RDA - a reality deconstruction algorithm would seek to reverse engineer the process of reification and create a real world in a sandbox. Such an algorithm would be able to create simulations with data about the real world. Observation data would be used to create the environment; ledger data would be used to populate the simulation with actors. Theoretically, there will be multiple versions of reality one can run to act on the world, measure outcomes and train the algorithm over time to give better and better predictions.

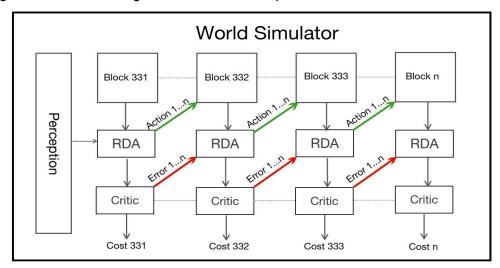


Fig. 3

Since a world simulator is a critical element for the system to exist, it could play the 'metabolism' role and maintain the life of an artificial intelligence.

4 Perception Module

The perception module of the system would have the task to estimate the state of the world. A person's everyday life requires an immense amount of knowledge about the world. People use the five senses to acquire it. To act intelligently a system would need a module with a similar function. It would need to acquire knowledge of the world, by extracting patterns from raw data. This capability is known as machine learning. One of its techniques is called deep learning and allows computer systems to use data for training and improve over time.

4.1 Deep learning

Deep learning is believed to be the only viable approach to the difficult task of representing real-world environments. Deep neural networks are already successfully applied to build object classifiers - the technical solution to the "typification" process. For example, a deep learning system can learn what an image of a fig fruit is by deconstructing it into smaller objects like corners, contours, edges etc. The first hidden layer would identify edges by comparing the brightness of neighboring pixels. The second layer would represent corners and contours as a collection of edges. Collections of contours and corners would then describe the third layer as objects. There will be a layer to describe every element of the model. The system is called "deep" because it can have a very large number of layers. And "learning" because it receives individual pixels and outputs object identity - it learns what a fig fruit is. To build the last module of the system, deep learning algorithms would be enhanced by dedicated hardware.

4.2 Hardware

Much of the current progress in the field of artificial intelligence was possible due to the exponential growth of computing power. Machine learning algorithms developed back in 70' and 80' proved to give impressive results when performed on very large training sets and dedicated hardware in the recent years. Even though the theoretical framework was in place, the field development depended on the hardware to catch up. An extrapolation of the process would mean that a major breakthrough in the field of artificial intelligence would require a major breakthrough in hardware design.

Moore's law [16] on which current progress relied upon is about to reach a major milestone in the field of classical digital computing by 2021. It would be physically impossible to pack smaller transistors to build more and more powerful

computers. Semiconductor industry intents to address the issue with an approach called "3D Power Scaling" [14]. It would be an attempt to print a 'skyscraper' of transistors into silicon. Back in the 50', stacking up vacuum tubes didn't power the next computer revolution [18]. It was the transistor that represented the much needed break-through. Perhaps, different approaches in hardware design need to be considered for the field to further develop. One of those is called quantum computing and promises different, better and faster computers.

4.3 Quantum Computing

Quantum Computing merges two great scientific revolutions of the 20th century: computer science and quantum physics. Quantum physics is the theoretical basis of the transistor, the laser, and other technologies which enabled the computing revolution. But on the algorithmic level, today's computing machinery still operates on "classical" binary logic. Quantum computing is the design of hardware and software that replaces Boolean logic by quantum law at the algorithmic level. For certain computations such as optimization, sampling, search or quantum simulation this promises dramatic speedups [15]. Whereas classical digital computing requires that the data be encoded into binary digits (bits), each of which is always in one of two definite states (0 or 1), quantum computation uses quantum bits, which can be in superpositions of states. This, together with gubits' ability to share a quantum state called entanglement, should enable the quantum computer to essentially perform many calculations at once, rather than in sequence like a traditional machine. And the number of such calculations should, in theory, double for each additional qubit, leading to an exponential speed-up. Quantum computing could be the technology to make sure an exponential growth of computing power would sustain the development of an artificial general intelligence in the foreseeable future.

5 Conclusion

Over the past years, important progress was made in the field of artificial intelligence. Which, could foster a hope that a human level intelligence could be developed in the lifetime of this generation. An artificial general intelligence could span human civilization to horizons never seen before. But, as with any powerful technology, it could also lead to civilization collapse. To avoid the latter, a safety-first approach is proposed. It would be a network of machine learning agents to learn what people value and a cryptographic ledger to make sure the system upholds that. Data about real world and the social constructs people created over generations would help to build the missing link of an artificial general intelligence - a world simulator. Quantum computing would provide the hardware such a system would need to run on. Once the system is up and

running and a thriving partnership between artificial intelligence and humanity is established, a star exploration era could begin, an era that transcends convenience.

5.1 References

- [1] Man-Computer Symbiosis* J. C. R. LICKLIDER 1960
- [2] A Path to Al | Yann LeCun, director of Al at Facebook, Beneficial Al Conference Asilomar, https://www.youtube.com/watch?v=bub58oYJTm0 2017
- [3] Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016.
- [4] Cooperative Reinforcement Inverse Learning https://arxiv.org/abs/1606.03137
- $\label{eq:continuous} \mbox{[5] Deep reinforcement learning from human preferences},$

https://arxiv.org/abs/1706.03741

- [6] Implications of AI for the Economy and Society | Jeffrey Sachs, Beneficial AI Conference Asilomar https://www.youtube.com/watch?v=Cyfl 8ucZPA, 2017
- [7] A counseling case adapted from Level 1 in Narrative Practice Training handout, INT UK 2011
- [8] Narrative Means to Therapeutic Ends, Michael White and David Epston 1990.
- [9] Superintelligence: Paths, Dangers, Strategies by Nick Bostrom, 2014
- [10] Mastering Bitcoin, Andreas M. Antonopoulos, 2014
- [11] Dyna, an integrated architecture for learning, planning, and reacting, Rich Sutton 1991
- [12] The Social Construction of Reality: A Treatise in the Sociology of Knowledge Peter L. Berger and Thomas Luckmann. 1966
- [13] Narrative Therapy: The Social Construction of Preferred Realities, by Gene Combs and Jill Freedman. 1996
- [14] International Technology Roadmap for Semiconductors (ITRS)
- https://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs/_2015
- [15] Quantum Computing and Deep Learning
- http://www.datasciencecentral.com/profiles/blogs/quantum-computing-and-deep-learning-how-soon-how-fast
- [16] Moore's Law by Gordon Moore, 1965
- [17] Sapiens: A Brief History of Humankind by Yuval Noah Harari 2014
- [18] The Dream Machine: J.C.R. Licklider and the Revolution That Made Computing Personal 2001