iSegoria[1]: a path to artificial general intelligence

Ion Dronic
dronic@isegoria.com
www.isegoria.com

Abstract

To build a safe artificial general intelligence there is a need for value alignment, a process to make sure the intelligent system comprehends what people value and why. For that matter, a counseling network is proposed for development. It would create a process for value discovery; use a cryptographic ledger to aggregate and store data; create an economic mechanism to incentivize everyone's participation and make sure the benefits of such a system are largely shared. Decentralized network architecture would propose a mechanism to control, influence and mitigate the existential risk such a system would pose to humanity. Gathered data and the social construction of reality would propose a solution to the artificial general intelligence missing link – the world simulator. Deep learning and neural networks would estimate the state of the world. It is uncertain that such a system can be built, but it would be worth trying.

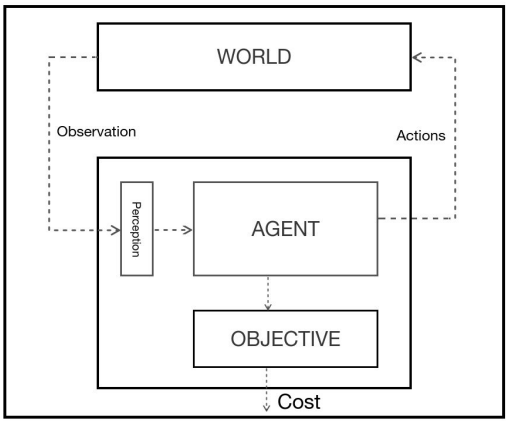
[1] isegoria - equality of all in freedom of speech (ancient greek)

1. Artificial General Intelligence

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. Licklider[2] 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. Voices[2] in the AI community suggest that an intelligent artificial system would emulate the human brain and would be composed of 3 main modules:

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

A thought experiment would explain how the system would work. 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. In 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.



An important aspect that should not pass unnoticed is the order of module development. If the agent module is developed first and the system's objective is misaligned, humans are in great danger. An artificial intelligence that would achieve a human level of intelligence would quickly transcend it [9]. 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 artificial general intelligence would behave the same, but it's better to be safe

[2] **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 also known as the founder of DARPA.

than sorry. In order to mitigate any of the doomsday scenarios, we should focus on developing the objective module first.

2. Objective Module

The objective 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. To achieve the task we propose an objective module composed of two submodules: value and control.

2.1 Value Submodule

The value module would make sure an artificial general intelligence would uphold the same values as people. There are several ideas under development that address the "value" 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 a person values by observing its behavior.

A typical counseling case would highlight some concerns this approach might raise. 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. If we would allow an intelligent system to learn what people value by observing their behavior we might end up with an agent that would learn that shouting at children is an acceptable behavior and this is what people want. It is a great challenge to learn what people value, but an opportunity as well. We could reinvent moral thinking and incorporate the knowledge that is often ignored or forgotten.

It was Jacques Derrida[3] who pioneered the idea that 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. Words are relational and are always based on the distinction with that which it is not – "injustice" only has significance in relation to "justice", distinguishing "despair" depends on an appreciation of "hope" etc. The invisible side is what Michael White [7] calls 'the absent but implicit'; that which is on the other side – and on which this description depends – is the absent but implicit. We would need a mechanism to learn people's day-to-day problems to figure out what are the values hidden behind the problems. Once the mechanism is established, we could

[3] Jacques Derrida was a French philosopher best known for developing a form of semiotic analysis known as deconstruction.

aggregate what people value on a global scale. To accomplish such a task we would propose a **counseling network** for development.

2.2 Counseling Network

To make sure an artificial general intelligence system upholds people's values, we would need a process of value discovery. 'Luckily', there are 300 million people we could learn from. As of March 2017 [8], there are 300 million people living with depression, according to the World Health Organization. We would apply the "absent but implicit" practice to learn what are the values hidden behind people's depression. The thing that is in jeopardy, would be what people value – the absent but implicit.

Personalized mental health care for 300 million people isn't an easy task and scaling the existing practices in a traditional manner won't suffice. That's why we would propose several innovations to address the scalability issue:

- **boot-camps** - a five-week training that would recruit new counselors among the people that went through counseling themselves.
- **intelligent assistant** - a chatbot that would make use of artificial narrow intelligence to speed up counselors' learning curve.
- **supervision** - the education process would continue on the network. A professional/pool of professionals would review every session that happens on the network. They would give feedback; share knowledge; point out a direction for further exploring.

The knowledge behind the innovations is called narrative approach to counseling. It seeks to be a respectful, non-blaming approach to counseling and community work that centers people as the experts in their own lives. It views problems as separate from people and assumes people have many skills, competencies, beliefs, values, commitments and abilities that will assist them to reduce the influence of the problems they deal with. However, all that knowledge lives in the stories people don't tell about themselves and it is a narrative approach to counseling that engages people in sharing their life stories; reflect and help them live the kind of life they prefer.

2.3 Boot-camps

Narrative counseling practice is seen rather as a social than psychological process. It focuses on exceptions and success, instead of problems and failure. That's why one does not need to be expert in disorders and dysfunctions to practice it. A narrative practice called "outsider witness practice" could be employed to imagine a new, scalable way to grow the existing community of counselors.

In the narrative, a person's sense of identity is seen as a social achievement: 'authenticity' is achieved when other people (outsider witnesses)[4] recognize and acknowledge their preferred identity claims. For that matter, Michael White [7] used to maintain registers with people who went through the process and are willing to help others. The network would benefit from such registers and seek to recruit counselors from the "outsider witness" community. The

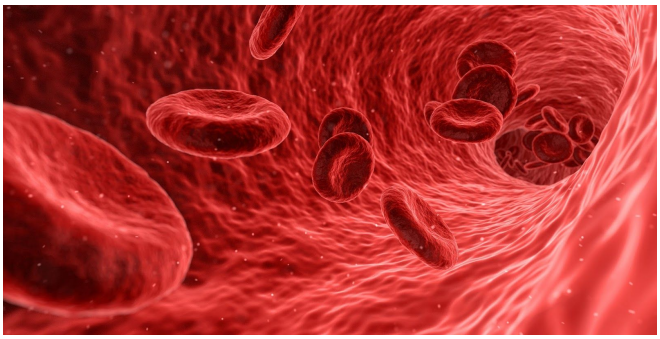
[4] An implementation of the "outsider witness" concept is the social media "like" button. Which, is nothing else but a mechanism to "authenticate" a person's preferred identity claims (look, food, idea, action etc)

network would then organize counseling boot camps – a 3-level training in narrative practice (5 weeks) and recruit new counselors from people that went through counseling themselves.

2.4 Intelligent Assistant

An intelligent assistant would employ artificial narrow intelligence to speed up counselors' learning curve. Industry examples of such systems 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. **Anni** (Artificial Narrow Narrative Intelligence) is the proposal for an artificial narrow intelligence agent trained in a narrative approach to counseling. Some counseling practices Anni could help out with: the story metaphor in counseling; externalisation; exceptions; de-centred practice; re-authoring lives; question structures; norms and power; wider contexts; outsider witnessing; writing letters and documents; re-membering conversations; revision; absent but implicit; trauma; maps; power, context and discourse; expressions of failure etc.

2.5 Control Submodule



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. Humans' vascular system would be the inspiration for such a module. Human blood is a stream of cells that carries oxygen to the brain. Which, is used to perform chemical reactions needed for brain to operate. A stream of cryptographic tokens would be the "blood" of the system. They would have a similar function - to carry the keys to decrypt counseling transcripts. If the blood stream stops the organism dies. If people would stop using the network the artificial intelligence would also cease to exist. Human body uses bone marrow to produce **blood cells**. We would use blockchain technology to produce **cryptographic tokens**.

2.6 Blockchain

On the network, data is organized in blocks. A block is a collection of counseling transcripts. The underlying technology is called blockchain (in a way it's a chain of blocks linked to each other). Data is stored in a distributed fashion where every scaffolding node holds a local copy of the data. This approach enables features that current record keeping systems are lacking: immutability, transparency, security and privacy.

The artificial intelligence system would be conditioned to train with the data stored on the network. The agent would need to pick one of the following scenarios to access it: rent, seize or bypass.

Scenario 1: Rent. The best way to do it would be to rent data from people. The agent would have to figure out some sort of economic activity to be able to pay for the decryption keys.

Scenario 2: Seize. 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 AI 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.

Scenario 3: Bypass. The agent can try to employ quantum 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. That's why the network won't keep "the wolf and the sheep on the same machine". The blockchain database with the transcripts would be located on the scaffolder's nodes of the network. The intelligent agent would run its code on the counselor's nodes. It can try to use the network to remotely decrypt the transcripts, but it would need to do it on thousands of nodes simultaneously (due to decentralized blockchain database architecture), which would require high bandwidth and an insane amount of energy.

The network would notice the attempt and launch countermeasures. An example of such a measure would be something called a "network fork" [10]. It would be 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 network data and the ability to run simulations. 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.

2.7 Cryptographic Token

Blockchain technology creates a powerful incentive mechanism for everyone to participate – a cryptographic token. It is a mechanism to anonymously prove ownership of an asset. It could be money, data or something else. The token would have a dual purpose in the system. Short-term, the it would consume and reward counseling services. Long-term, the token would become a mean to reward people for renting their personal data. We propose **NAR** to be the token name and **n** – the symbol**.** To foster mass adoption of the token and associated counseling services the network would launch its own crypto-exchange service. It would allow buying, selling and trading the token. The ability to gain economic benefits would represent the perfect "excuse" for people to join the network. It could become an important mechanism to mitigate the fear of stigma.

2.8 Network Architecture

A blockchain counseling network would have a peer-to-peer architecture on top of the Internet. While nodes in a p2p network are equal, they may take on different roles depending on the functionality they are supporting[10]. A node would be a collection of modules: routing, messaging, reviewing, the chatbot and the blockchain database.

N - the routing module would enable nodes to communicate on the network.
M - the messaging module would enable two peers to message each other.
C - the chatbot or the intelligent assistant would facilitate the counseling process.
R - the review module would enable a professional to give feedback on a counseling session.
B - the blockchain module, a database that would maintain a history with all session transcripts.
All nodes discover and maintain connections to peers and propagate session transcripts. However, different module combinations would assume different roles on the network.

1. **User** - (N, M) any person that seeks help can run such a node. The node would need the messaging module to communicate with other peers and a network module to connect to the network.
2. **Counselor** - (N, M, C) a person that went through a counseling training and is willing to help other peers can run such a node. The node would need an additional chatbot module.
3. **Scaffolder** - (N, R, B) a person that has years of experience in counseling can run such a node. The node would need a module to review session transcripts and a blockchain module to store the data.

2.9 Consensus Algorithm

The network would need to figure out a democratic and transparent way of determining the counselor to serve a help request and a practitioner to review a session transcript. Such a process is similar to what other decentralized networks call “consensus algorithms” (Proof-of-Work, Proof-of-Stake) or mechanisms to determine what network node would process the next batch of transactions/transcripts. Since the network seeks to **build** a global ledger of knowledge, “**scaffolding**” would be the name for such an algorithm. SCA (Scaffolding Consensus Algorithm) would be composed of two reputational algorithms:

- CRA: Counselor Reputation Algorithm - to match a person that is seeking help (**User**) with a person that struggled with a similar problem and went through a counseling training (**Counselor**). It would calculate a compatibility score based on several factors: the type of problem people are dealing with, the language the two peers are using, counselor’s boot-camp rating score, review score, counselor responsiveness and seniority. The result of each match would be a counseling transcript. Which, would be signed with 3 private keys belonging to the user, the counselor and the chatbot.
- SRA: Scaffolder Reputation Algorithm - to pick a pool of professionals (**Scaffolders**) to give feedback on the last 50 session transcripts that happen on the network. It would calculate two additional components: the amount of data professional counselors contributed when joined the network (work done before the inception of the network is the only way to determine a counselor’s experience); a random component to increase or decrease the score (to prevent centralizing of requests). The result of the match would be a block of counseling transcripts. Which, would be signed by a professional.

It is important to acknowledge that the ownership of data is shared between the actors that participated in its creation: user, counselor, chatbot, scaffolder. They would produce the cryptographic stream of tokens or the blood of the system.

2.10 Use Case

A network use case:

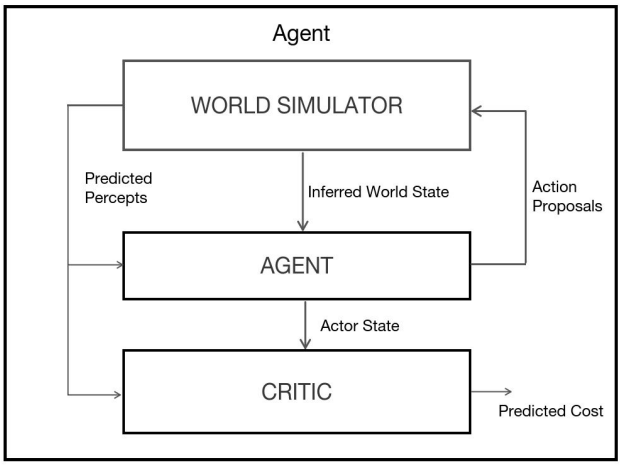
1. A person downloads the User App.
2. The chatbot will interview the user and reward him with tokens.
3. The user broadcasts a problem and an amount of tokens to get help.
4. CRA algorithm would match the user with a counselor.
5. A counseling session starts. The counselor listens, the chatbot facilitates the process.
6. The counseling session ends. A transcript of the session is broadcasted to all network nodes. Once received, the transcript is added to a pool of unverified transcripts.
7. SRA algorithm picks a scaffolder to review the oldest 50 transcripts from the pool.
8. A review/supervision process takes place. If a session transcript met the narrative counseling quality requirements it is marked as verified.
9. All the verified transcripts are added to a block. This block is added to the blockchain database and the system issues an amount of tokens to reward the scaffolding node/nodes – the mechanism that would ensure the supply of tokens on the network.
10. All nodes that acted as counselors and got their transcripts verified receive the tokens associated with each session.

3. Agent Module

The agent would be the module that would generate a sequence of actions to act on the world and measure the outcome of its actions. 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 (absent but implicit). 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. An intelligent agent would have to incorporate a world simulator, an actor and a critic function to be able to predict outcomes of its actions.



Thought Experiment 3: An 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 **inferred state** of the world (world simulator) to **predict** what product the company is hiring for. It would use a **critic** function to judge if the prediction is good or not. 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 industry developed ways to build most of the agent components with one exception: The world simulator - the missing link of an artificial general intelligence (LeCun 2017) [2]. The network would further develop the concept of Dyna and propose a solution to the world simulator problem.

3.2 World Simulator

“The Social Construction of Reality” a treatise in sociology by Berger and Luckmann [12] would guide the research towards building a world simulator. They propose the idea that people together construct their realities as they live them. A thought experiment proposed by Combs & Freedman [13] would explain the concept better.

“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 (1966, p. 60) puts it “An institutional world…is experienced as an objective reality.” They propose several processes that are important in the way any social

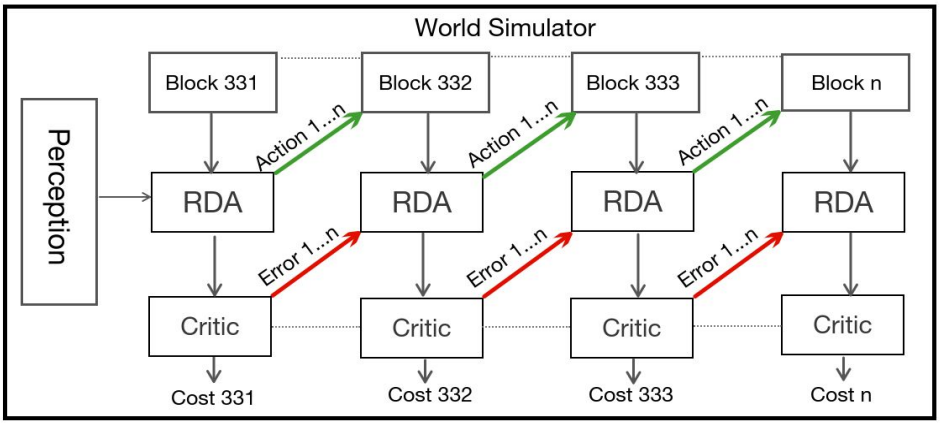
group constructs and maintains its knowledge concerning “reality”: typification, institutionalization and legitimation. Reification is the term to encompass the overall process of which the other three are parts.

- **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 Berger and Luckmann use to refer to those processes that give legitimacy to the institutions and typifications of a particular society.
- **Reification**, the result of the combined processes of typification, institutionalization, and legitimation – the result of creating a reality, it becomes taken-for-granted.

To build a virtual, simulated world there is a need to translate the mentioned processes into code. Because of this, a reality deconstruction algorithm is proposed for development.

3.3 Reality Deconstruction Algorithm

RDA - a reality deconstruction algorithm would be network’s proposal to translate the social construction of reality into code. The algorithm would seek to reverse engineer the process of reification. Deep neural networks [8] are already successfully applied to build object classifiers. And since there is a technical solution to the “typification” process, the other two processes could be also emulated with enough research and development. The RDA algorithm would use data acquired by observation to train and data stored on the network to populate the simulation with actors. As mentioned before, the network data is stored in blocks linked to each other. A block is a collection of session transcripts. A transcript is an honest, unfiltered perception of a person’s reality. One could think of a counseling session as a person’s snapshot of reality, description of his world, people and problems a person faces on a day-to-day basis. That’s when the network data would leverage its full potential - it would become a mechanism for people to gain economic benefits by renting out their perception of reality. A person’s snapshot of reality introduced into a World Simulator is a suitable environment to “try things” and measure outcomes.



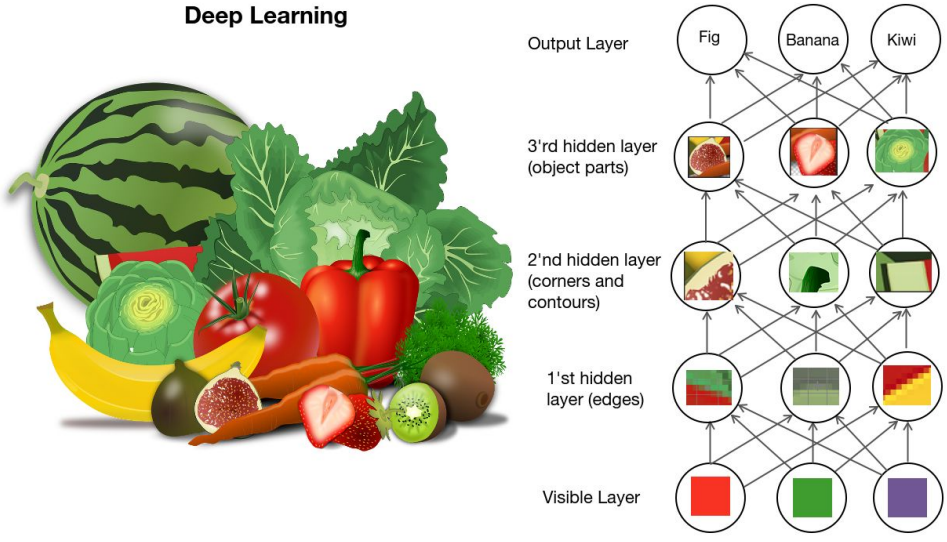
The picture above illustrates a world simulator. The agent would have a predefined sequence of actions to run; measure the expected error using the critic function; adjust the action proposals to optimize the critic and train itself to produce better and better outcomes. The longer chain becomes over time, the more accurate predictions could be. The decentralized nature of the network would allow real-time parallel simulations on millions of devices. In a sense, it would mimic the parallel computing human brain is so good at – only 20 watts of energy powers the most complex system in the universe.

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 learn 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 a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones"[3]. It is believed to be the only viable approach to the difficult task of representing real-world environments.



For example, a deep learning system can learn what a image of a fig is by deconstructing that image in smaller objects like corners, contours, edges etc. There will be layer to describe every element of the model. The first hidden layer would easily identify edges by comparing the brightness of neighboring pixels. The second layer would represent corners and contours as

collection of edges. Collections of contours and corners would than describe the third layer as objects and so on. Such a system is called “deep”, because it can have a very large number of layers. It is called “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 we would use deep learning and dedicated hardware.

5. Conclusion

Over the past years important progress was made in the field of artificial intelligence. Which, could foster the hope that a human level intelligence could be developed in the lifetime of this generation. We believe that artificial intelligence could span human civilization to horizons never seen before. As with any powerful technologies it could lead either to a thriving partnership either to a civilization collapse. To avoid the latter, we propose a safety first approach. It would be a counseling network that would learn what people value and make sure the system upholds that. The network’s decentralized architecture would make sure there is a mechanism to control; govern and mitigate the existential risk such a system would pose to humanity. Gathered data and the social construction of reality would help building the missing link of an artificial general intelligence - a world simulator. Deep learning and dedicated hardware would estimate the world’s state. Once the system is up and running, it could help out solving most of the problems current society is dealing with: poverty, hunger, climate change etc. And then, an era that transcends convenience could begin.

5.1 Acknowledgements

The writing of this paper is a 2-year joint effort of multiple institutions and people. We would like to thank them for their contribution.

- **Innovation Norway** – for funding the research and development of this paper.
- **Psiterra,** romanian narrative therapy association – for guidance and training in narrative counseling.
- **Draper University** – for providing a two-month entrepreneurship program in Silicon Valley, California USA.
- **The Future of Life Institute** – for providing a cutting edge insight on the state of current artificial intelligence development by organizing the Beneficial AI conference in Asilomar, California January 2017.

5.2 References

[1] - Man-Computer Symbiosis* J. C. R. LICKLIDER
http://worrydream.com/refs/Licklider%20-%20Man-Computer%20Symbiosis.pdf
[2] - A Path to AI | Yann LeCun, director of AI at Facebook, Beneficial AI Conference Asilomar 5-7 January 2017, https://www.youtube.com/watch?v=bub58oYJTm0
[3] - “Deep Learning” by Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016.
[4] - Cooperative Reinforcement Inverse Learning https://arxiv.org/abs/1606.03137
[5] - Concrete Problems in AI Safety, https://arxiv.org/abs/1606.06565
[6] - Level 1 in Narrative Practice Training handout, INT UK 2011
[7] - Narrative Means to Therapeutic Ends, Michael White and David Epston 1990.
[8] - World Health Organization Report, March 2017
http://www.who.int/mediacentre/news/releases/2017/world-health-day/en/
[9] - “Superintelligence: Paths, Dangers, Strategies” by Nick Bostrom
[10] - Mastering Bitcoin, Andreas M. Antonopoulos
[11] - Dyna, an integrated architecture for learning, planning, and reacting, by Rich Sutton.
http://dl.acm.org/citation.cfm?id=122377
[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)