*First human upload as AI Nanny*

Working draft

*Alexey Turchin*

Digital Immortality Now

Foundation Science for Life Extension

alexeiturchin@gmail.com

**Abstract**: As there are no visible ways to create safe self-improving superintelligence, but it is looming, we probably need temporary ways to prevent its creation. The only way to prevent it, is to create special AI, which is able to control and monitor all places in the world. The idea has been suggested by Goertzel in form of AI Nanny, but his Nanny is still superintelligent and not easy to control, as was shown by Bensinger at al. We explore here the ways to create the safest and simplest form of AI, which may work as AI Nanny. Such AI system will be enough to solve most problems, which we expect the AI will solve, including control of robotics, acceleration of the medical research, but will present less risk, as it will be less different from humans. As AI police, it will work as operation system for most computers, producing world surveillance system, which will be able to envision and stop any potential terrorists and bad actors in advance. As uploading technology is lagging, and neuromorphic AI is intrinsically dangerous, the most plausible way to human-based AI Nanny is either functional model of the human mind or a Narrow-AI empowered group of people, probably inside governmental structure like NSA.

**Disclaimer**: Some ideas may sound morally and politically controversial. They don’t represent author’s political point of view.

**Highlights**:

* It is impossible to make safe powerful AI, but the only instrument to prevent its creation is another AI.
* AI must exist and non-exist simultaneously.
* AI should not be separated agent from humans in order to escape alignment problem.
* Several possible solutions exist: human upload as a first AI, human augmentation and human empowered organization.
* The most probable solution is that NSA will get strategic decisive advantage via Narrow AI empowered human organization combined with nuclear superpower resources.

[1.Introduction 2](#_Toc1490521)

[2. AI alignment is intrinsically difficult task so better to escape it all 3](#_Toc1490522)

[2.2. AI alignment as two-agent problem 3](#_Toc1490523)

[2.3. AI Nanny 4](#_Toc1490524)

[3. Human upload as a form of intrinsically Safe AI 4](#_Toc1490525)

[3.1. Attractive properties for a human upload as Safe AI 4](#_Toc1490526)

[3.2. Potential problems with human upload as a Safe AI 6](#_Toc1490527)

[3.3. Human upload is intrinsically safe against intelligence explosion, but still have many failure modes 8](#_Toc1490528)

[4. Different ways to the safe and perfect human upload 9](#_Toc1490529)

[HLAI: Functional model of human brain as replacement of full brain scans 9](#_Toc1490530)

[Human mind secret is in human training dataset 11](#_Toc1490531)

[AI-nafts and AI-president 11](#_Toc1490532)

[Human-like observer core 13](#_Toc1490533)

[Data driven very large AI 13](#_Toc1490534)

[BCI and Neuralink; net of self-improving individuals 13](#_Toc1490535)

[How to accelerate creation of the uploads? 14](#_Toc1490536)

[Narrow way between human upload and neuromorphic AI 14](#_Toc1490537)

[Combining human uploads and NSI-AI approaches 15](#_Toc1490538)

[Narrow AI for human upload scaffolding 16](#_Toc1490539)

[Combining human uploads and NSI-AI approaches 16](#_Toc1490540)

[7. Conclusion. 17](#_Toc1490541)

[References 17](#_Toc1490542)

# 1.Introduction

This article is pessimistic. It assumes that there is no way to create safe benevolent self-improving superintelligence, and the only way to escape its creation is implementation of some form of limited AI, which will work as a Global AI Nanny, controlling and preventing dangerous AIs appearance as wells as other global risks.

The idea of AI Nanny was suggested by Goertzel. I explored its levels of realization in (Turchin & Denkenberger, 2017a), where 4 levels were suggested: based on global surveillance, neural net’s based global brain, net of narrow AI as AI police and mild form of superintelligence. However, AI Nanny need not to be superintelligent itself, or all the same control problems will appear.

In this article, we will explore the ways how idea to non-superintelligent AI Nanny which could provide non-perfect but workable solution for the global AI safety problem, which could be probably done by some combination non-perfect human uploads, human augmentation, organizational superintelligence and Narrow AI.

The idea of using human uploads or human augmentation as some form of limited and safer AI has been suggested by several authors (Hanson, 2016), (Kelley & Waser, 2016) (Eth, 2017), (Sotala & Valpola, 2012), (Salamon & Shulman, 2011), (Mennen, 2016), but it has some theoretical and practical difficulties.

To solve AI Safety we will apply the theory complex problems soling created by (Altshuller, 1999) in Section 2, explore the theoretical advantages of human uploads in Section 3, and practical ways to the upload creation in the section 4, the ways to reach decisive advantage via Narrow AI in section 5.

# 2. AI alignment is intrinsically difficult task so better to escape it all

## 2.2. AI alignment as two-agent problem

The problem of “AI alignment”, that is ensuring that AI has the value system which is aligned with value system of its human creators is the problem of relation of two agents, which is aggravated by the fact that these two agents have very different capabilities and properties. This have many subproblems, including, but not limited to:

* Values in humans and in AI are presented in different forms, and we don’t know how they are presented in humans.
* Human value system consists of many values and following literary just one command, could contradict many other hidden values
* Commands and values which are safe and clear on one level of capabilities, could become dangerous on higher level of capabilities, or on larger scales.
* Human values are integrated in human common sense, and without human common sense, they could be dangerous.
* Not all human values should be installed in AI: we don’t want that AI will have sex drive to humans.
* AI’s values should not be just parallel to human one’s, AI should be under human control
* It is not clear which subset of human values should be instilled into AI.

The AI alignment is also called AI Control problem, and it is a problem how weaker agent will control infinitely more powerful agent. We could see that all above listed problems are arising from not easy relation of *two* agents, and if there will be only one agent, the situation will be much simpler. There are several ideas how to escape “two agents problem”:

* Non-agential AI, like “Oracle AI” or “Tool AI”.
* Human augmentation as a way to AI.
* Human organization as AI.
* Limiting robotic general AI to very predictable set of behavior, like in case of self-driving cars.

Full escape of agential AI is not easy, as robots will appear (and robotic brain is probably agential) and as agential optimization is more effective and could appear inside Oracle AI or Tool AI. If we escape two agents problem, we become closer to the solution where AI exist and non-exist simultaneously.

## 2.3. AI Nanny

Goertzel suggested to create mild form of AI, which will have most benefits of AI, but escape risks of unpredictable Strong AI, and called it *AI Nanny*. He suggested that AI Nanny will be as advanced above human as humans are above chimps. Thus, *AI Nanny* will be superintelligent anyway, and, as Bensinger showed most of the problems to control AI Nanny will be the same as needed to control any other Strong AI. (However, the explanation in his article is smaller than expected, and possible counterarguments are better presented in original Goertzel article.)

The main problems with AI Nanny is that:

* We don’t know how to create AI.
* We don’t know how to install human values in AI, if we have it.
* We don’t know how to control even mild superintelligence.
* We don’t know how to reach global coordination needed to AI Nanny implementation. all over the world without starting all-out war for planet modification.
* We don’t know how prevent AI Nanny from self-improving, wrong conclusions from its goal system, and most other risks associated with full-blown superintelligence.

In this article, we suggest the way to construct AI Nanny, combining human uploading, side-loading, and AI control system, as well as use of market forces to deliver it almost everywhere.

# 3. Human upload as a form of intrinsically Safe AI

## 3.1. Attractive properties for a human upload as Safe AI

Only handful of papers explore benefits of human uploads for AI safety (Mennen, 2016)(Eth, 2017)(Salamon & Shulman, 2011).

Human upload is having many attractive properties:

* Has human values.
* Not easy to self-improve.
* Its intelligence is not above human (and could be increased in measurable and reversible way by speeding up or running several copies).
* We will be able to understand, what is going on with it.
* It is able to understand context correctly.
* We could understand what it thinks, and it can’t hide its thought stream (as it is presented in the form of internal dialog).
* If human consciousness is important for cognition and AI safety, it could be modeled or installed into the upload.
* We know typical human failure modes and know how to deal with them.
* Uploads are error prone, as human mind has evolved in the way that it is able to compensate some neurons deaths and even small strokes.
* Even if biological humans will extinct, human upload will survive.

We could also choose actual set of values by uploading a human with known set of values. It could be a human with extensive training in moral philosophy, programming, AI safety theory and any other relevant disciplines. He may prepare to such uploading for years, and be copy-positive (that is, positive that many of his copies will exist and will be often rebooted).

As human upload is very large and messy computer program, it is not easy to improve the upload, or its improvement can’t gain exponential growth. We may run it quickly, or use analogues of some stimulating drugs, but such improvement will be marginal, and it is better, as it thus will be controllable and predictable, and could be turn off. If we run human upload 1000 times quicker, we could get roughly log(1000) times increase in performance (Turchin & Denkenberger, 2017b).

What is more important, it is difficult to improve the upload from inside, as internal complexity of its messy code is greater than the size of its memory. Expected size of a human upload is 1015 bytes, and its working memory is the same as in humans.

Human upload may run on higher speed and will be overstimulated, but its general level of intelligence will be on the same level as of humans, so its logic and actions could be matched and understand by a group of people.

Even if everything goes wrong, and only a human upload(s) survives, it will be still continued of humanity.

## 3.2. Potential problems with human upload as a Safe AI

1. As Bostrom and others wrote, *creation of human upload may not prevent later creation of non-human superintelligence* (Bostrom, 2014), and even help in it by providing needed insight and instruments. ––Reply: In our case, of AI Nanny based in human upload, it will be not a problem, as we will use the upload based AI to prevent superintelligence creation.

2*. Messy values presentation in human uploads may produce strange and dangerous behavior.* We know many example of bad humans, or good humans affected by short emotion, like anger, who did a lot of crimes. Also, humans are known to have many undesired traits. Do we want that AI will have sex drive and want to have sex with human beings? What about depression, suicide ideation? Hitler was a human too. – Reply: bad behavior is not default mode for humans, and it could be fixed by Narrow AI scaffolding discussed below.

*2a. Small variations in human mind upload will produce dystopian outcomes.* 90 per cent of a human value is not human value at all, but something uncanny similar. Reply: as human upload could be tested, this could be compensated.

*2b. The goals of neuromorphic AI are unknown for the outside viewer and even to itself.* Reply: Human mind is able to understand its goals and rules and such understanding is more important than blind following the rules, which create all sorts of problems like paperclip maximizers. Human minds also could see rules as laws. Human mind is also controlled by external reward, provided through emotions and created by its older part.

*2c. Uploads could have hidden failure models*.The canonic (but fictional) example is neural net designed to recognized tanks but in fact learned to recognized day and night (gwern, 2011). Neural nets have different adversarial examples than humans (for whom such adversarial examples are optical and other illusions). But after it neural nets become able to demonstrate superhuman performance, while symbolic and mathematical AI’s can’t do it – in part because of messy nature of surrounding reality. As result, messy AIs appears more robust than pure symbolic AI, and are able to increase quality of their performance arbitrary high.

*2d. Some human values and traits are incompatible with being Nanny or useful upload*.

* Most humans prefer not to be turn off after task finishing
* Humans have sex drive, which we don’t expect from robots
* Humans tend to abuse power, become power hungry, corrupted, pleasure addicted.
* Humans have many known cognitive biases
* Human minds have limited capacity in complex tasks.
* Humans have idea of revolt and treason.
* Their mood fluctuates.
* The may feel like in a prison, or like a slave while being an ems.
* They could be affected by mental disorders, and being without a body and without friends could affect their state of mind.
* Human mind may become maniac about anything, and behave as small scale paperclipper (hoarders, psychopaths, sexual maniacs and fetishists). Such maniacs mostly limited by the resources.
* Humans are affected by all sorts of biases.

3. *Human uploads may evolve unpredictably using some forms of evolutionary strategies.* However, we will prevent such use of uploads by AI Nanny. After we create and test the AI Nanny, its internal content will be black boxed by anti-reverse engineering techniques and by other boxing solution which we explore in “Catching treacherous turn”. It will still be able to increase its performance by running on higher speed.

*4. Small tweaks of a human upload could make it completely misbalanced*, like some analogous of simulating of drug’s effect.

5. *Goertzel thinks that “nannification” of the upload will be difficult*. “If one is talking about, say, an AGI built via closely emulating human brain architecture, in which the designers have relatively weak understanding of the AGI system’s representations and dynamics, then the ‘nannification is hard’ problem might be more serious”. However, we have much better knowledge of human brain when of any other exiting AI (as no other exist), and could extensively test it.

*6. Current technology is lagging in human modelling*. We still don’t know a lot about the ways the brain function. Our modeling of neural system of animals stuck with nematode worm for 30 years, despite that all connections of its small brain of 137 neurons are known. Expected size of the first uploads will be so large, that they will need biggest supercomputers, like used in Blue Brain project, and so they could not be used widely. We could replace direct uploading with functional brain modeling, side loading and learning, as it is discussed in the next section.

*7. Accelerating of WBE research will help neuromorphic AIs,* which are regarded dangerous. It is regarded bad because neuromorphic systems are messy, and we can’t prove anything mathematically about their future behavior. However, the current success of neural nets is in some sense triumph of neuromorphic AI, and thus additional progress in this direction becomes less dangerous, as many neuromorphic messy things already exist.

We may also mention that this objection is not about the safety of the uploads, but about its possible effects on AI evolution in general. However, it put into the line with other considerations, like that the total refuse to create AGI would increase risks of bioweapons and that for non-human AIs will be more difficult to understand the context and this will probability of the “paperclip maximizer” outcome. All roads into the future are dangerous, so information that one particular road is dangerous is not very strong.

Another idea which needs farther considerations is that *Attempts to create AGI using neuromorphic approach will create mind architecture, similar to human mind.* Basically, if we want AGI to see, it needs eyes, and if we want it to understand human language, it should have language processing, and if we want to use as robot brain, it has to have motor cortex. We also should note that industry demand for natural language processing and robotic brains plus the use of neural net technologies is already steering direction into something, which has general functions of human brain (but not universal task solver). We called it “robotic brain”.

8. *Ethical concerns about enslaving a human being* in the computer simulation.

HLAI works and is felt from inside as an ordinary human brain. <https://intelligence.org/files/SS11Workshop.pdf>

## 3.3. Human upload is intrinsically safe against intelligence explosion, but still have many failure modes

As upload is intrinsically safe AI design (doesn’t kill everybody immediately after first start), we could solve most of its problems as they appear during testing. It is its big advantage, as every error in other AI designs will be fatal.

There are several ways how to create superintelligence starting from uploads, but they are mostly do not produce explosive growth:

* Mennen showed the way of uploads self-improving via genetic algos (Mennen, 2016)
* Brains region stimulation
* Virtual organizations form many brains
* Speed acceleration
* Uploads coalescence (Sotala & Valpola, 2012).
* Neuromorphic AI inspired by uploads. This especially true for connectome based uploads, as they could help to solve secret of intelligence, but less for functional model, which only produce what was input in them.

This again reminds us classical TRIZ problems solving situation: there should be a human upload and should not be human upload.

# 4. Different ways to the safe and perfect human upload

As we said above in 2.3, we need a human upload, but simultaneously it should not be an upload, because such upload will carry all bad human traits with it. Also, brain scanning is delaying, but neuromorphic architectures in the form of neural nets and robot brains are winning.

In this section, we will explore several ways to such “safe” upload. Uncertainty and speed of changes in the field is so large, that we can’t here take responsibility for choosing just one path. We should explore different paths and later choose the best.

## HLAI: Functional model of human brain as replacement of full brain scans

One approach to safer human upload is creation of idealized model of human mind on functional level using all available instruments. We would call it HLAI – human-like artificial intelligence.

Reason for HLAI creation are following:

1. Direct human brain uploading is catastrophically lagging compared to other AI system, which makes all the discussion purely theoretical. The main problem is that we still not able to model brains of even of smallest and well scanned animals, like *C.Elegance*, and also we still don’t know many simplest things about human brain, like why we sleep and see dreams, how we encode memory, there is the center of consciousness etc.

2. Several large brain modeling projects could produce meaningful results in 2020s, like Blue Brain and Allen Brain. But modeling of a brain of an old random person will not be a Safe AI (but could provide needed insights about the nature of intelligence for non-human projects.)

3. Most current successes in AI are using neural nets and are somehow similar to human brain (however, most successful convolutional nets are different from spiking neural nets, similar to human brain).

Thus, if we want that human uploads come first, we need to replace long and messy brain scanning with higher-level process of brain modelling, which will be in line with industry.

Such brain modelling will use several techniques:

1. **Functional model** means that we create a system from many blocks, which are connected the same way as blocks inside the human brain. Each block is a black box, so we need to model its inputs and outputs, but not what is going on in it on neuronal level. The idea is based on that we can already model functionality of some brain blocks. So, we could work on each block independently and attract to it all newly appearing neurological and psychological data.

We also could use all recent successes in machine learning to do what we want inside each black box. If we model a hippocampus, it should not work inside as a human hippocampus, but could use any available ML systems. (And such approach was used for a chip for replacement of the hippocampus.)

We could also use scans of brain parts as soon as we get them.

These ideas are in line with current economic development, which create demand for robotic brains for home-robots and self-driving cars. Such robotic brains must have almost the same functionality as human brains – situation awareness, ability to understand language, apply ethics, obey commands safely. Thus, robotic companies will try to make such brains anyway.

This non-self-improving safer AI is analogous to non-self-replicating safer molecular manufacturing [Phoenix, Chris, and Eric Drexler. "Safe exponential manufacturing." Nanotechnology 15.8 (2004): 869.].

The main problem here is that too simple and effective model of human brain will be exactly the type of universal AI which we try not to create.

2. **Sideloading**. We could extensively test the system, tweak and see if its behavior is similar to the human mind.

Intrinsic messiness of human brain may be replaced by cryptographic protection of internal content of different black boxes, from which the system will be constructed.

Some general model of human mind is created using functional modelling and all other tricks, and then it is tweaked until its behavior become similar to the behavior of given person. (But it is not clear, how to install in it large set of personal memories without long repetition of the education process.)

3. **Learning**. As neural nets are very general, they could be taught almost everything, and they could be trained to mimic human behavior as a black box. The same way neural nets are now trained to drive self-driving cars: they are fed with hundreds of thousands of hours of videorecording of actual human driving and become able to demonstrate roughly human levels performance in complex street driving.

## Human mind secret is in human training dataset

This idea is that human minds are rather blank neural net-based learners, and what makes us humans is extensive training on our cultural dataset. There are several pieces of evidence for it:

1. Feral children. Humans grown by wolves are behave like wolves and can’t learn language (if they are older than 7 years) and adapt to human culture in other ways.
2. Monkeys can learn some human-like behavior like smoking and gest language, and express ideas that they are humans.
3. Anatomical humans existed for hundreds of thousand years, but was not able to start civilization, probably because they didn’t create proper training dataset for intellectual activity.
4. Neural nets could be trained to behave like human car driver just by observations of correct behavior.
5. People with significant brain abnormalities are able to produce most of human behavior, that is, to be human is very robust to architectural changes of the mind.

From the point of AI safety, it means that if we take a blank neural net with approximately the same functional architecture as a human brain, and train it on typical human dataset (may be by raising it as a child, or just demonstrating videos about human life), it will learn how to be human and even would think that it is a human.

This means that it is especially important to prepare ethical dataset (ref – Shegurin), because if a mind is trained in a cruel environment, it will replicate cruel behavior.

## AI-nafts and AI-president

Another approach is go all-in in a very large project for preparing a perfect human being and scanning this person, so s/he will become the first upload. In 2018 Brain Preservation Prize was awarded to Nectome cryo-chemical preservation technology, and this could be first step to uploading.

The great thing of idea of human upload as first and only AI is that this plan consists of several clear steps without obvious intractable difficulties: preparing the needed person, developing brain scanning technologies, developing large enough supercomputer.

To create Safe AI, we don’t want to model just a random person, who will probably an old man, donated his brain for scientific research, but we should prefer to have a scan of very special person, who will be able and also want to work inside computer as a global AI Nanny.

It was suggested before that “safety-conscious human researchers could be “uploaded” into computing hardware and, as uploads, will solve problems relating to AGI safety before the arrival of AGI” – but in this case uploads are used as AI researchers, not as AI Nanny. (Salamon & Shulman, 2011), and also these ideas were discussed in (Salamon & Muehlhauser, 2011). They also suggested “WBE safety monitors” and also that WBE could evolve by copying with small changes.

We also would like not just one “AI-president”, but some form of “AI-council” of different minds. (We should note that humans are adapted to be governed by one person, or to govern a country, that is, humans have clear set of cultural stereotypes, how to do govern or be governed.)

In fact, we may need only one such human, who may solve all his task in many copies and in accelerated timeline.

What will be discussed next may rise some ethical concerns, but this path is imaginable:

* A special person is educated in science, moral philosophy and practice plus many other disciplines, as well as tested for performance.
* S/he is copy-positive (agree to be copied), self-less, etc.
* S/he agree to sacrifice his biological life to be the first scanned man and to work as AI in solving global problems.
* S/he will be euthanized and his-her brain will be scanned as soon as possible and after some testing s/he is used as the first and only AI, and works as AI Nanny as well as on some other important tasks, like curing aging and preventing other global risks.

No such person is currently known to for the authors, as people are not ideal (and if a person looks like an ideal he could be a psychopath, so finding him-her will be a “winner curse”).

It looks like current western society will not accept such sacrifice. However, it could be framed as first journey in unknown, the same way as astronauts risked their lives to explore space. The reward for such person will be immortality, some form of global power and eternal fame (or hate). The risks are losing of the personal identity, failure of scanning process to recreate his personality or decision not use him as AI Nanny if s-he fails subsequent testing. May be non-western country, like China, will allow such a procedure.

## Human-like observer core

J.Kelley and Waser suggested creating human-like observer core inside AI (Independent Core Observer Model Cognitive Architecture, ICON) which will have simplified model of human emotions and consciousness and ensure AI safety (Kelley & Waser, 2016).

The reason for this idea is that human mind is also hierarchical, and at the top is its conscious content, its values, its decisions and experience of emotions, while a lot of complex machinery like visual procession is unobservable and non-important. For a human upload, we may not need most of these hidden parts of the mind, but only correctly working top part.

However, it is not clear, if we can create this top-model without exact modeling of the frontal lobe (and some other parts) where language processing and complex abstract thinking is happening. In that case, creating observer core will be not much simpler than creating the whole functional model of the brain or creation of AI.

## Data driven very large AI

There is another potentially safe AI architecture – data driven AI, example of which is a library or Google search engine. Basically, it is a type of Oracle AI, which gets its power from the search of very large database. So, we can get answers, but it doesn’t use much optimization power.

## BCI and Neuralink; net of self-improving individuals

There is another way to a human upload, that is self-improving and self-upgrading individual. Great side is that there is no need to uploading, and there is no “two-agents’ problem”: that is relations between creator and its creation.

Elon Musk took this approach with his Neuralink.

**Global surveillance system with humans in the loop.** Another way to escape difficulties of human uploading is not upload humans at all, but to use some specific form of human augmentation. This specific form of human augmentation is not something like exocortex which is typically imagined, but:

* global surveillance system.
* powerful analytic tools, like *Palantir*.
* effective system of task distribution between group of people.
* individual human’s minds augmentation via stimulators, training and some other instruments.

## How to accelerate creation of the uploads?

1. HLAI approach: human brain architecture is trained of human dataset in safe and controlled environment, until it will behave almost as human (but better)
2. Neural robes approach
3. Full brain connectome scan of a dead person

## Narrow way between human upload and neuromorphic AI

Neuromorphic AI is an AI which mimics some aspects of the human mind and inspired by brain science data. It is regarded as potentially dangerous as it could have get a “secret of intelligence” from human brains studies, but not human values. As a result, it will quickly outperform humans because of the hardware overhang and it will evolve into a non-aligned superintelligence, – at least according to narrative promoted by MIRI. However, may be lose some value by ignoring neuromorphic approaches and could get more information if we look into them deeper?

There are following reasons for it:

1. It seems that like “marginal danger” of the perusing of the neuromorphic approach is declining, because humans are doing this anyway, and local ignoring of these trend is not changing global perspective.
2. The biggest impact of the neuromorphic ideas has already happened in form of neural nets AI architecture. Also, there is no much hidden ideas which we don’t know, we just need time and hardware to implement them and test (ref Christiano) many neuromorphic ideas are in the air – Markram; Hawking, On intelligence.
3. There are different possible ways to the neuromorphic AI.

Any upload is not a real “human mind copy”: there is always some form of the simplification. However, on some level of abstraction it could become a neuromorphic AI - on functional level.

Types of the neuromorphic AIs:

1. *Functional model of the human brain*. Example: LIDA architecture. There will be huge demand for such brains which solve typically human problems in typically human environments, first of all, for the development of home robots. It must control a body, navigate in human created environment, understand common language commends – and do most thing that a typical human maid will do, like preparing tea. This type of AI is most obvious type of KANSI, where existing modules are collected in something which is behave as human brain and model around a human mind. The treat here is that an AI can get some kind of high-level understanding of itself, and this high-level reasoning abilities will be where the point of dangerous self-improving starts (Like in the *West World* TV series).
2. *“Cortical column model based*” – solving some low-level principle of the intelligence and the parallelize it. (“On intelligence”).
3. *Neural net trained to mimic human behavior*. If we fed the neural net with thousands of hours of a person’s activity, it presumably could behave in the same way. Ok, it will not actually think, but how we know that humans are *actually* thinking – maybe they are just trained to behave as if they thinking? Some experiments showed that neural nets are able to perform correct arithmetic operations just by training on the handwritten examples.

## Combining human uploads and NSI-AI approaches

We explored two approaches where AI exists and doesn’t exist simultaneously: Human uploads based AI and secret intelligence organization of a superpower based AI (NSI-AI). The first is more technically difficult, but better, and NSI-AI is more realistic in near term, but seems to be less moral, as secret services are known not be not aligned with general human values, and will reach global domination probably via illegal covert operation including false flags attacks and terrorism.

It seems that this two approaches should be united, and some ideas exist:

1. *Simple HLAI works as data crunching assistances inside NSI-AI*. Maybe they be will not be full AGI, but robotic brains with some natural language understanding, which could help accelerate search, communication, filtering, idea generation etc. They are rather small addition inside communication lines which increase personal human capabilities and collaborative capabilities of the group.
2. *HLAI as part of world model simulation.* To win in world domination game one needs effective world model. If such finite world model is created, and is accurate, the systems like *AlphaZero* could find winning strategies, where small actions will have large consequences. The main problem here is principle uncertainty of the world which makes such planning difficult, but one-step planning (where plan is not long chain of action of some mastermind but is corrected after consequences of every action is possible (Sotala, 2017)).
3. *First KANSI or global brain is created inside NSI,* using some principles of human brain architecture, and works as an Oracle AI and this provides the NSI with significant knowable advantage over other data crunching organizations.
4. *Secret human uploading project has been long underway somewhere in DARPA,* and first upload will be a member of the military (may be not killed, but chosen from deadly ill group). DARPA provided a lot of funding to neurointerfaces and neuromodeling (Hatmaker, 2017), (Strickland, 2017).
5. *NSI-AI may steel other project design and run it on much stronger computers.*

## Narrow AI for human upload scaffolding

We could use narrow AI to compensate all failures modes of human upload and to steer its behavior in needed way. Large boxing system mainly to prevent AI self-improvement we described in the “Catching treacherous turn”. But most of ideas from it will be not needed in case of human upload.

The functions of Narrow AI scaffold around human upload:

* Record thought stream (for safety and backup)
* Regulate emotions level (Upload should have some levels of emotions, or it will be non-human, but many human failure modes are associated with strong emotions, like attacks of anger.)
* Reward useful behavior.
* Detecting mental illnesses
* Prevent unauthorized access and copying.
* Connection with Tools AI and databases

## Combining human uploads and NSI-AI approaches

We explored two approaches where AI exists and doesn’t exist simultaneously: Human uploads based AI and secret intelligence organization of a superpower based AI (NSI-AI). The first is more technically difficult, but better, and NSI-AI is more realistic in near term, but seems to be less moral, as secret services are known not be not aligned with general human values, and will reach global domination probably via illegal covert operation including false flags attacks and terrorism.

It seems that this two approaches should be united, and some ideas exist:

1. *Simple HLAI works as data crunching assistances inside NSI-AI*. Maybe they be will not be full AGI, but robotic brains with some natural language understanding, which could help accelerate search, communication, filtering, idea generation etc. They are rather small addition inside communication lines which increase personal human capabilities and collaborative capabilities of the group.
2. *HLAI as part of world model simulation.* To win in world domination game one needs effective world model. If such finite world model is created, and is accurate, the systems like *AlphaZero* could find winning strategies, where small actions will have large consequences. The main problem here is principle uncertainty of the world which makes such planning difficult, but one-step planning (where plan is not long chain of action of some mastermind but is corrected after consequences of every action is possible (Sotala, 2017)).
3. *First KANSI or global brain is created inside NSI,* using some principles of human brain architecture, and works as an Oracle AI and this provides the NSI with significant knowable advantage over other data crunching organizations.
4. *Secret human uploading project has been long underway somewhere in DARPA,* and first upload will be a member of the military (may be not killed, but chosen from deadly ill group). DARPA provided a lot of funding to neurointerfaces and neuromodeling (Hatmaker, 2017), (Strickland, 2017).

*NSI-AI may steel other project design and run it on much stronger computers*

# 7. Conclusion.

If we have a lot of time, the way to AGI through human uploads may have significant safety advantage, as uncontrollable intelligence explosion is unlikely and result is intrinsically safe. The main question is how to but this time.

We could reach it either by slowing down other ways to AGI via AI police and other drastic control measures, or via significant acceleration of human brain scanning and neural probe technologies.

# References

AlexMennen. (2017). Existential risk from AI without an intelligence explosion. Retrieved from http://lesswrong.com/lw/p28/existential\_risk\_from\_ai\_without\_an\_intelligence/

Altshuller, G. S. (1999). *The innovation algorithm: TRIZ, systematic innovation and technical creativity*. Technical Innovation Center, Inc.

Bostrom, N. (2014). *Superintelligence*. Oxford: Oxford University Press.

Christiano, P. (2016). Prosaic AI alignment. Retrieved from https://ai-alignment.com/prosaic-ai-control-b959644d79c2

Eth, D. (2017). The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes. *Informatica*, *41*(4).

gwern. (2011, September 20). The Neural Net Tank Urban Legend - Gwern.net. Retrieved March 27, 2018, from https://www.gwern.net/Tanks

Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press.

Hatmaker, T. (2017, July 10). DARPA awards $65 million to develop the perfect, tiny two-way brain-computer interface. Retrieved March 26, 2018, from http://social.techcrunch.com/2017/07/10/darpa-nesd-grants-paradromics/

Kelley, D. J., & Waser, M. R. (2016). Implementing a Seed Safe/Moral Motivational System with the Independent Core Observer Model (ICOM). *Procedia Computer Science*, *88*, 125–130. https://doi.org/10.1016/j.procs.2016.07.415

Krakovna, V. (2015, November 30). Risks from general artificial intelligence without an intelligence explosion. Retrieved March 25, 2018, from https://vkrakovna.wordpress.com/2015/11/29/ai-risk-without-an-intelligence-explosion/

Mennen, A. (2016). Superintelligence via whole brain emulation. Retrieved March 25, 2018, from /posts/jMKZKc2GiFGegRXvN/superintelligence-via-whole-brain-emulation

Salamon, A., & Muehlhauser, L. (2011). Singularity Summit 2011 Workshop Report. Retrieved from https://intelligence.org/files/SS11Workshop.pdf

Salamon, A., & Shulman, C. (2011). Whole Brain Emulation, as a platform for creating safe AGI. MIRI. Retrieved from http://slideplayer.com/slide/6079210/

Sotala, K. (2016). Decisive Strategic Advantage without a Hard Takeoff. Retrieved from http://kajsotala.fi/2016/04/decisive-strategic-advantage-without-a-hard-takeoff/#comments

Sotala, K. (2017). How feasible is the rapid development of artificial superintelligence? *Physica Scripta*, *92*(11). https://doi.org/10.1088/1402-4896/aa90e8

Sotala, K., & Valpola, H. (2012). Coalescing minds: brain uploading-related group mind scenarios. *International Journal of Machine Consciousness*, *4*(01), 293–312.

Strickland, E. (2017, July 10). DARPA Wants Brain Implants That Record From 1 Million Neurons. Retrieved March 26, 2018, from https://spectrum.ieee.org/the-human-os/biomedical/devices/darpa-wants-brain-implants-that-record-from-1-million-neurons

Turchin, A. (2017). Messaging future AI. Retrieved from https://goo.gl/YArqki

Turchin, A., & Denkenberger, D. (2017a). *Global Solutions of the AI Safety Problem*. manuscript.

Turchin, A., & Denkenberger, D. (2017b). *Levels of self-improvment of AI*.

<http://spectrum.ieee.org/computing/software/what-intelligent-machines-need-to-learn-from-the-neocortex>