*Narrow AI Nanny:*

 *Reaching Strategic Advantage via Narrow AI to Prevent Creation of the Dangerous Superintelligence*

Alexey Turchin

Digital Immortality Now

Foundation Science for Life Extension

alexeiturchin@gmail.com

**Abstract**: As there are no currently obvious ways to create safe self-improving superintelligence, but its emergence is looming, we probably need temporary ways to prevent its creation. The only way to prevent it is to create a special type of AI that is able to control and monitor the entire world. The idea has been suggested by Goertzel in the form of an AI Nanny, but his Nanny is still superintelligent, and is not easy to control. We explore here ways to create the safest and simplest form of AI which may work as an AI Nanny, that is, a global surveillance state powered by a Narrow AI, or AI Police. A similar but more limited system has already been implemented in China for the prevention of ordinary crime. AI police will be able to predict the actions of and stop potential terrorists and bad actors in advance. Implementation of such AI police will probably consist of two steps: first, a strategic decisive advantage via Narrow AI created by an intelligence services of a nuclear superpower, and then ubiquitous control over potentially dangerous agents which could create unauthorized artificial general intelligence which could evolve into Superintelligence.

**Highlights**:

* Narrow AI may be used to achieve a decisive strategic advantage (DSA) and acquire global power.
* The most probable route to DSA via Narrow AI is the creation of Narrow AI by the secret service of a nuclear superpower.
* The most probable places for its creation are the US National Security Agency or the Chinese Government.
* Narrow AI may be used to create a Global AI Police for global surveillance, able to prevent the creation of dangerous AIs and most other existential risks.
* This solution is dangerous but realistic.

[1. Introduction 2](#_Toc520298241)

[2. The main contradiction of the AI safety problem: AI must simultaneously exist and not exist 3](#_Toc520298242)

[3. Decisive strategic advantage via Narrow AI 4](#_Toc520298243)

[3.1. Non-self-improving AI can obtain a decisive advantage 4](#_Toc520298244)

[3.2. Narrow AI is used to create non-AI world-dominating technology 5](#_Toc520298245)

[3.3. Types of Narrow AI which may be used for obtaining a DSA 7](#_Toc520298246)

[3.4. The knowability of a decisive advantage 8](#_Toc520298247)

[4. AI-empowered reconnaissance organization of a nuclear superpower is the most probable place of origin of a Narrow AI DSA 9](#_Toc520298248)

[4.1. Advantages of a secret Narrow AI program inside the government 9](#_Toc520298249)

[4.2 Existing governmental and intelligence Narrow AI projects according to open sources 10](#_Toc520298250)

[4.3. Who is winning the Narrow AI race? 11](#_Toc520298251)

[5. Plan of implementation of AI police via Narrow AI advantage 11](#_Toc520298252)

[5.1. Steps of implementing of AI safety via Narrow AI DSA 11](#_Toc520298253)

[5.2. Predictive AI Police based on Narrow AI: what and how to control 12](#_Toc520298254)

[6. Obstacles and dangers 13](#_Toc520298255)

[6.1. Catastrophic risks 13](#_Toc520298256)

[6.2. Mafia-state, corruption, and the use of the governmental AI by private individuals 13](#_Toc520298257)

[Conclusion. Riding the wave of the AI revolution to a safer world 13](#_Toc520298258)

# 1. Introduction

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

The idea of AI Nanny was first suggested by Goertzel (Goertzel, 2012); we have previously explored its levels of realization (Turchin & Denkenberger, 2017a). An AI Nanny does not itself need to be a superintelligence, as if it is, all the same control problems will appear again (Muehlhauser & Salamon, 2012).

In this article, we will explore ways to create a non-superintelligent AI Nanny via Narrow AI. Doing so involves addressing two questions: First, how to achieve a decisive strategic advantage (DSA) via Narrow AI, and second, how to use such a system to achieve a level of effective global control sufficient to prevent the creation of superintelligent AI. In the sister article, we look at the next level of AI Nanny, based on human uploads, which currently seems a more remote possibility, but which may become possible after implementation of a Narrow AI Nanny (Turchin, 2017).

The idea of achieving strategic advantage via AI before the creation of the superintelligence was suggested by Sotala (Sotala, 2018), who called it a “Major strategic advantage” as opposed to a “Decisive strategic advantage”, which is overwhelmingly stronger, but requires superintelligence. A similar line of thought was presented by Alex Mennen (Mennen, 2017).

Historically, there are several examples where an advantage in Narrow AI has been important. The most famous is the case is breaking of German cipher *Enigma* via electro-mechanical “cryptographic bombe” constructed by Alan Turing, which automatically generate and tested hypothesis about code (Welchman, 1982). It was an overwhelmingly more complex computing system than any other during WW2, which gave the Allies informational domination over the Axis powers. A more recent, but also more elusive, example is the case of Cambridge Analytica, which supposedly used its data-crunching advantage to contribute to the result of the 2016 US presidential elections (Cottrell, 2018). Another example is the use of sophisticated cyberweapons like Stuxnet to disarm an enemy (Kushner, 2013).

The Chinese government’s facial recognition and human ranking system is a possible example not of a Narrow AI advantage, but of “global AI police”, which create informational dominance over all independent agents; however, any totalitarian power which worth its name had effective instruments for such informational domination even before computers, like Stasi in the former East Germany.

To solve AI safety we will apply the theory of complex problem solving created by Altshuller (1999) in Section 2; discuss ways to reach a decisive advantage via Narrow AI in section 3; and, in section 4, examine ways to use Narrow AI to effectively monitor and prevent creation of unauthorized self-improving AI. In section 5 we will look at ways to safely develop AI Police based on an advantage in Narrow AI, and in section 6 we will examine potential failure modes.

# 2. The main contradiction of the AI safety problem: AI must simultaneously exist and not exist

It is becoming widely accepted that sufficiently advanced AI may be global catastrophic risk, especially if it becomes superintelligent in the process of recursive self-improvement (Bostrom, 2014; Yudkowsky, 2008). It has also been suggested that we should apply engineering standards of safety to the creation of AI (Yampolsky & Fox, 2013).

Engineering safety demands that the creation of the unpredictably explosive system whose safety cannot be proved (Yampolskiy, 2016) or incrementally tested should be prevented. For instance, no one wants a nuclear reactor with unpredictable chain reaction; even in a nuclear bomb, the chain reaction should be predictable. Hence, if to really apply engineering safety to the AI, there is only one way to do it:

*Do not create artificial general intelligence (AGI).*

However, we can’t prevent creation of AGIs by other agents as there is no central global authority and ability to monitor all AI labs and individuals. In addition, the probability of global cooperation is small because of the ongoing AI arms race between US and China (Ding, 2018; Perez, 2017).

Moreover, if we postpone the creation of AGI, we could succumb to other global catastrophic risk, like biological risks (Millett & Snyder-Beattie, 2017; Turchin, Green, & Denkenberger, 2017) as only AI-powered global control may be sufficient to effectively prevent them. We need powerful AI to prevent all other risks.

In the words of problem solving method *TRIZ* (Altshuller, 1999), the core contradiction of the AI problem is following:

*AGI must exist and non-exist simultaneously*.

What does it mean for AI to “exist and non-exist simultaneously”? Several ways to limit the capabilities of AI so it can’t be regarded as “fully existing” have been suggested:

1. *No agency*. In this case, AI does not exist as an agent separate from humans, so there is no alignment problem. For example, AI as a human augmentation, as envisioned in Musk’s Neuralink (Templeton, 2017).
2. *No “artificial*” *component*. AI is not created *de novo*, but is somehow connected with humans, perhaps via human uploading (Hanson, 2016). We will look more at this case in another article, “Human upload as AI Nanny”.
3. *No “general intelligence*”. The problem-solving ability of this AI arises not from its wit, but because of its access to large amounts of data and other resources. It is also Narrow AI, not a universal AGI. This is the approach we will explore in the current article.

# 3. Decisive strategic advantage via Narrow AI

## 3.1. Non-self-improving AI can obtain a decisive advantage

Recently Sotala (2016), Christiano (2016), Mennen (2017), and Krakovna (2015) have explored the idea that AI may have a DSA even without the capacity for self-improvement. Mennen wrote about following conditions for the strategic advantage of non-self-improving AI:

1. *World-taking capability outperforming self-improving capabilities,* that is, “AIs are better at taking over the world than they are at programming AIs” (Mennen, 2017).He suggests later that, hypothetically, AI will be better than humans at some form of engineering. Sotala opined that, “for the AI to acquire a DSA, its level in some offensive capability must overcome humanity’s defensive capabilities” (Sotala, 2016).
2. *Self-restriction in self-improvement*. “An AI that is capable of producing a more capable AI may refrain from doing so if it is unable to solve the AI alignment problem for itself” (Mennen, 2017). We have previously discussed some potential difficulties for any self-improving AI (Turchin & Denkenberger, 2017b). Mennen suggests that AI’s advantage in that case will be less marked, so boxing may be more workable, and the AI is more likely to fail in its takeover attempt.
3. *Alignment of non-self-improving AI is simpler*. “AI alignment would be easier for AIs that do not undergo an intelligence explosion” (Mennen, 2017), as it will be a) easy to monitor its goals, b) less of a difference will be observed between our goals and the AI’s interpretation of them. This dichotomy was also explored by Maxwell (2017).
4. *AI must obtain a DSA not only over humans, but over other AIs*, as well as other nation-states. The need to have advantage over other AIs depends on the number and relative difference between AIs producing teams. We have looked at the nature of AI arms races in an earlier paper (Turchin & Denkenberger, 2017a). A smaller advantage will produce a slower ascension, and thus a multipolar outcome will be likely.

Sotala added a distinction between the major strategic advantage provided by Narrow AI and that of DSA by superintelligent AI (Sotala, 2018). Most of what we will describe below falls in the first category. The smaller the advantage, the riskier and more uncertain its implementation, and the process of the implementation could be more violent.

In the next subsections we will explore how Narrow AI may be used to obtain a DSA.

## 3.2. Narrow AI is used to create non-AI world-dominating technology

Narrow AI may be implemented in several ways to obtain a DSA, and for a real DSA, these implementations should be combined. However, any DSA will temporary, and may be in place for no more than one year.

**Nuclear war-winning strategy**. Narrow AI systems could empower strategic planners with the ability to actually win a nuclear war with very little collateral damage or risk of global consequences. That is, they could calculate a route to a credible first strike capability. For example, if nuclear strategy could be successfully formalized, like the game Go, the country with the more powerful AI would win. There are several ways in which such nuclear superiority could win using AI:

* *Strategic dominance*. Create a detailed world model which could then be played in the same way as a board game. This is most straightforward way, but it is less likely, as creation of a perfect model is unlikely without AGI and is difficult in the chaotic “real world”.
* *Informational dominance*. The ability to learn much more information about the enemy, e.g. the location of all its nuclear weapons and the codes to disable them. Such informational dominance may be used to disarm the enemy forces; it may also include learning all state secrets of the enemy with guaranteed preservation of their own secrets.
* *Identify small actions with large consequences.* This category includes actions such as blackmail of the enemy’s leaders and the use of cryptoweapons and false flags to corner the enemy. This approach will probably will work if combined with strategic dominance.
* *Dominance in manufacturing*. New manufacturing technology enables cheaper and deadlier missiles and other military hardware like drones and large quantity of them. This especially apply to invisible weapons for first strike, like stealth cruise missiles.
* *Deploy cyberweapons inside the enemy’s nuclear control chains.* Something like an advanced form of a computer virus embedded in the nuclear control and warning systems.

Dominance in nuclear war does not necessarily mean that actual war will happen, but such dominance could be used to force the enemy to capitulate and agree to a certain type of inspections. However, a credible demonstration of the disarming capability may be needed to motivate compliance.

**New technology which helps to produce other types of weapons.**

* *Biological weapons.* Advances in computer empowered bioengineering could produce targeted bioweapons. It may be not worthwhile to list all possible hazards which an unethical agent could use in a quest for global domination if the agent has access to superior biotechnology with science-fiction-level capabilities.
* *Nanotechnology*. Molecular manufacturing will allow the creation of new types of invisible self-replicating weapons, much more destructive then nukes.

**Cyberweapons**, that is, weapons which consist of computer programs and mostly affect other programs.

* Hidden switches in the enemy’s infrastructure.
* The ability to sever communication inside an opposing military.
* Full computerization of the army from the bottom to the top (De Spiegeleire, Maas, & Sweijs, 2017).
* Large drone swarms, like the slaughterbots from a famous video (Oberhaus, 2017) or their manufacturing capabilities (Turchin & Denkenberger, 2018a).
* Financial instruments.
* Human-influencing capabilities (effective social manipulation like targeted adds and fake facts).

## 3.3. Types of Narrow AI which may be used for obtaining a DSA

There are several hypothetical ways how narrow AI could reach DSA.

One is **Data-driven AIs**: systems whose main power comes from access to the large amounts of data, which compensate for their limited or narrow “pure” intelligence. This includes subcategory of “*Big brothers”.* This category includes systems of criminal analysis like Palantir (recently mocked in the Senate as “Stanford Analytica” (Midler, 2018)), which unite mass surveillance with the ability to crunch big data and find patterns. Another type is *World simulations.* World simulations may be created based on data collected about the world and its people to predict their behavior. The possessor of the better model of the world would win.

**Limited problem solvers** are systems which outperform humans within certain narrow fields which includes:

* “*Robotic minds*” with limited agency and natural language processing capabilities able to empower a robotic army, for example, as the brain of a drone swarm .
* *Cryptographic supremacy*. The case of Enigma shows the power of cryptographic supremacy over potential adversaries. Such supremacy might be enough to win WW3, as it will result in informational transparency for one side. Quantum computers could provide such supremacy via their ability to decipher codes (Preskill, 2012).
* *Expert systems as Narrow Oracles*, which could provide useful advice in some field, perhaps based on some machine learning-based advice-generating software.
* *Computer programs able to win strategic games.* Something like a strategic planner with playing abilities, e.g. Alpha Zero (Silver et al., 2017). Such a program may need either a hand-crafted world model or connection with the “world simulations” described in section 1.2. Such a system may be empowered by another system which able to formalize any real-world situation as a game.
* *Narrow AI in engineering* could dramatically increase the effectiveness of some form of weapons construction, for example, nuclear or biological weapons, nanotechnology, or robotics.

Narrow AI advantage may take also a form of **Narrow AI increasing the effectiveness of group intelligence**. *This could be Graphical collective thinking systems*, something like dynamic collectively edited roadmaps, wikis, or Palantir. One attempt to create such a platform was Arbital (Arbital, 2017). Christiano et al.’s “amplify and distill” project works on factored cognition, which will be a smartphone app which distributes different portions of cognitive tasks between teams (Ought, 2018). Also, it may take form of *AI-empowered personal search assistants*, maybe with a simple brain–computer interface or *Communication assistants,* which help to make conversation productive, record a conversation log and show relevant internet links. Finally, group intelligence may be aggregated via *large, self-improving organizations,* which implement all types of collective intelligence, hardware producing capabilities, money to hire the best talent, etc., like Google.

Sotala has discussed “minds coalescence” as a way to create more powerful minds (Sotala & Valpola, 2012). Danila Medvedev suggested that the use of a powerful collaborative information processing system, something between a Wikipedia, Evernote, and Mindmap, may significantly increase group intelligence. Similar ideas have been discussed by “Neuronet” enthusiasts like Luksha, where collective intelligence will be produced via brain implants (Mitin, 2014).

Superforecasting technology (Tetlock & Gardner, 2016) that aggregates predictions as well as prediction markets could be used to increase power of the “group brain”. In Soviet times this was known as “sharashka” (Kerber & Hardesty, 1996) – scientific lab consisted from imprisoned scientists, who were under government control and under pressure to make discoveries.

**Narrow AI able to reach “informational dominance” over all potential enemies**: in this situation, the enemy can’t have any secrets and all its actions are constantly monitored. This could be achieved via: sophisticated spyware in all computers; quantum computers for code breaking or some exotic quantum tech like quantum radar or quantum calculations using close time like curves; microscopic robots, as small as a grain of salt, which could be secretly implanted in the adversary’s headquarters.

## 3.4. The knowability of a decisive advantage

Even if one side reaches the level of decisive advantage which provides it with the opportunity to take over the world, it may not realize what it possesses if it doesn’t know the capabilities of other players, which could be made deliberately vague. For example, in the 1940s, the US had nuclear superiority, but the Soviet Union made vague claims in 1947 that the nuclear secret was no longer secret (Timerbaev, n.d.), thus creating uncertainty about its level of nuclear success.

To ensure a DSA, a rather invasive surveillance systems would need to be implemented first; in other words, the advantage must be reached first in informational domination, to guarantee knowledge of the capabilities of all opponents. This could be done via AI created inside an intelligence service.

A DSA provided by Narrow AI will probably require a combination of several of the Narrow AI types listed in section 3.3, and the only way to guarantee such dominance is the actual size of the project. The size will depend on resource investments, first of all, money, but also minds, and strategic coordination of all these projects into one workable system. It looks like only the US and China currently have the resources and determination needed for such a project.

If there is no knowable DSA, both sides may refrain from attacking each other. Armstrong et al. have created a model of the role of AI and mutual knowledge (Armstrong, Bostrom, & Shulman, 2016). Bostrom has also written about the topic in his article about AI openness (Bostrom, 2017).

A semi-stable solution consisting of two AIs may appear, as predicted by Lem (1959) and previously discussed by us (Turchin & Denkenberger, 2018b). Such a balance between two superpowers may work as a global AI Nanny, but much less effectively, as both sides may try to rush to develop superintelligent AI to obtain an insurmountable advantage.

Narrow AI provides a unique opportunity for knowable DSA. For example, the creators of cryptological bombe were not only able to break the codes of the enemy, but they probably know that they outperformed the code breaking technologies of the Axis, as the Axis didn’t mention the existence of their own code breaking and, more obviously, didn’t start to use harder codes, which they would have done if they had similar code-breaking technology. A Narrow AI-based DSA, based on “informational domination” creates a unique opportunity for an almost peaceful world takeover that also includes AI Police able to prevent the creation of unauthorized superintelligent AIs.

# 4. AI-empowered reconnaissance organization of a nuclear superpower is the most probable place of origin of a Narrow AI DSA

## 4.1. Advantages of a secret Narrow AI program inside the government

During discussions at MIRI (at the time, the Singularity Institute) in the 2000s, the idea that government and military structures would be interested in creating superintelligent AI was dismissed, because it was considered that the governments were too stupid to understand future AI capabilities, and thus creation of AI in a small private company was regarded more likely. But now it certainly not the case.

There are several reasons why a Narrow AI-driven decisive strategic advantage could be achieved inside the governmental structure of the large nuclear superpowers, and moreover, inside a secret intelligence and data crunching agency, similar to the National Security Agency (NSA) of the US. A nuclear superpower is already interested in world domination, or at least interested in preventing domination by other players. If geopolitics can be modeled as a strategic game, Narrow AI will help to achieve advantage in such game, as existing Narrow AIs demonstrate significantly superhuman abilities in winning in complex games, similar to the games for world dominance, like Go.

A nuclear superpower has almost unlimited money for secret AI project compared with startups and commercial corporations. Historically, the data-crunching capabilities of secret services have outperformed civilian applications. An AI of the same power as a civilian one but in the hands of a nuclear superpower could dramatically outperform the civilian AI. Military AI could leverage several non-AI advantages in the hands of the superpower: access to the nuclear weapons, large computational resources, nets of sensors, pools of big data, a large concentration of experienced researchers, and other secret state programs.

Such a secret government AI organization could take advantage of the openness in the field of AI, as it could absorb information about the advances of others, but would not be not legally obliged to share its own achievements. Thus, it would always outperform the current state of public knowledge. Governmental organizations have used this type of advantage before to dominate in cryptography.

## 4.2 Existing governmental and intelligence Narrow AI projects according to open sources

When we speak about Narrow AI inside a reconnaissance organization, we mean AI as a technology which increases the efficiency of data crunching within an organization which already has many advantages: very powerful instruments to collect data, money, access to secret technology, and attracts the best minds, as well as ability to educate and train them according to its standards.

The US NSA has been described as the world's largest single employer of [mathematicians](https://en.wikipedia.org/wiki/Mathematician) (and there are several other computer-related security agencies in the US) (Love, 2014). The NSA employs around 40 000 people (Rosenbach, 2013) and has a budget of around 10 billion USD. For comparison, Google employs 72 000 thousand people in 2016 (Statista, 2018).

NSA works on world simulations with humans (Faggella, 2013) and has vowed to use AI (B. Williams, 2017). Wired has reported that “MonsterMind, like the film version of Skynet, is a defense surveillance system that would instantly and autonomously neutralize foreign cyberattacks against the US, and could be used to launch retaliatory strikes as well” (Zetter, 2015). An interesting overview of governmental data crunching is presented in the article “The New Military-Industrial Complex of Big Data Psy-Ops” (Shaw, 2018). It was reported that the CIA runs 137 secret AI projects (Jena, 2017). However, it is useless to search open data about the most serious AI projects aimed at world domination, as such data will doubtless be secret.

An example of a Narrow AI system which could be implemented to achieve a DSA is *Palantir*, which was used for so-called “predictive policing technology” (Winston, 2018). Palantir is an instrument to search large databases about people and find hidden connections. Such a system also probably facilitates the collective intelligence of a group: conversation support Narrow AI may record and transcribe conversation on the fly, suggest supporting links, generate ideas for brainstorming and works as a mild Oracle AI in narrow domains. We don’t claim here that the Palantir is an instrument intended to take over the world, but that a Narrow AI providing a decisive strategic advantage may look much like it.

Another illustrative example of the Narrow AI systems we are speaking about is the Chinese *SenseTime*, which stores data describing hundreds of millions of human faces and is used for applications like the Chinese social credit system (Murphy, 2018).

## 4.3. Who is winning the Narrow AI race?

It looks like the US is losing the momentum to implement any possible strategic advantage in Narrow AI for political reasons: the conflict of the Trump administration with other branches of power; Snowden-type leaks resulting in public outcry; and the campaign against military AI collaboration with the government within Google (Archer, 2018). If this is the case, Chinese could take this advantage later, as their relationship with private organizations is more structured, political power is more centralized and ethical norms are different (Williams, 2018). There are several other powerful intelligence agencies of nuclear powers, like Russian or Israel, which could do it, though the probability is lower.

However, recent Narrow AI empowered election manipulation has happened not through direct action by governments but via a small chain of private companies (Facebook and Cambridge Analytica). This demonstrates that Narrow AI may be used to obtain global power via manipulation of elections.

In some sense, a world takeover using AI has already happened, if we count the efforts of Cambridge Analytica in the US election. But it is unlikely that Russian hackers combined with Russian intelligence services have the decisive strategic advantage in Narrow AI. What we observe looks like more of a reckless gamble based on a small temporary advantage.

# 5. Plan of implementation of AI police via Narrow AI advantage

## 5.1. Steps of implementing of AI safety via Narrow AI DSA

The plan is not what we recommend, but just the most logical way of action for a hypothetical “rational” agent. Basically, this plan consists of the following steps:

1. Gaining a knowable decisive advantage.
2. Implementing it for the world takeover.
3. Creating a global surveillance system (AI Police) that controls any possible sources of global risk, including biological risks, nuclear weapons and unauthorized research in AI.
4. Ban advanced AI research altogether or slowly advance it via some safe path.

While the plan is more or less straightforward, its implementation could be both dangerous and immoral. Its main danger is that the plan means starting a war against the whole world without an infinitely large advantage that could be ensured only via superintelligence. War is always violent and unpredictable. We have written previously about the dangers of military AI (Turchin & Denkenberger, 2018b).

There is nothing good about such a plan; it would be much better if all countries would instead peacefully contribute to the UN and form a “committee for prevention of global risks”. This is unlikely to happen now but may occur if an obvious small risk of a global catastrophe appears, such as an incoming asteroid or a dangerous pandemic. The problem of the creation of such a committee requires additional analysis into how to use the momentum of emerging global risks to help such a committee to form, become permanent, and act globally without exceptions. Even if such a committee were peacefully created, it would still need AI Police to monitor dangerous AI research.

## 5.2. Predictive AI Police based on Narrow AI: what and how to control

Even if world domination is reached using Narrow AI, such domination is not final solution, as the dominating side should be able to take care of all global problems, including climate change, global catastrophic risks and, first of all, the risks of the appearance of another, even more sophisticated or superintelligent AI which could be unfriendly.

We will call “AI Police” a hypothetical instrument which is able to prevent the appearance of dangerous AI research anywhere on the globe. There are two interconnected questions about AI police: what and how should be monitored?

Such a system should be able to identify researchers or companies involved in illegal AI research (assuming that the creation of superintelligent AI is banned). AI police instruments should be installed in every research center which presumably has such capabilities, and all such centers or researchers should be identified. Similar systems already was suggested to search for hackers (Brenton, 2018).

AI police may identify signs of potentially dangerous activity (like smoke as a sign of fire). Palantir was used in New Orleans for “predictive policing”, where potential criminals were identified via analysis of their social network activity and then monitored more closely (Winston, 2018).

Such an AI Police system will do all the same things that intelligence agencies are doing now; the main difference is that there will be no blind spots. The main problem is how to create such a system so it does not have a blind spot in its center, which often happens with overcentralized systems. Maybe such system could be created without centralization, based instead on ubiquitous transparency, or some type of net horizontal solution.

Many possible types of Narrow AI with a DSA, e.g. one based on informational domination via superiority of the information gathering and data crunching technology, could be directly transformed into AI Police. Other possible types, like a Narrow AI winner in the nuclear strategic game, could not be used for policing. In that case, additional solutions should be quickly invented.

# 6. Obstacles and dangers

## 6.1. Catastrophic risks

If one side wrongly estimated its advantage, the attempt to take over the world may result in world war. In addition, after a successful world takeover, a global totalitarian government, “Big Brother”, may be formed. Bostrom has described such an outcome as an existential risk (Bostrom, 2002). Such a world government may indulge in unlimited corruption and ultimately fail catastrophically. Attempts to fight such a global government may produce something another risk, like catastrophic terrorism.

If the “global government” fails to implement more advanced forms of AI, it may be not able to foresee future global risks; however, if it does try to implement advanced forms of AI, a new level of AI Control problems will appear. Such a world government may be not the best approach to solve it.

Not every attempt at global takeover via Narrow AI would necessarily be aimed at prevention of superintelligent AI. It is more likely to be motivated by some limited set of nationalistic or sectarian goals of the perpetrator, and thus, even after a successful takeover, the AI Safety problems will continue to be underestimated. However, as the power of Narrow AI will be obvious after such a takeover, control over other AI projects will then be implemented.

## 6.2. Mafia-state, corruption, and the use of the governmental AI by private individuals

While a bona fide national superpower could be imagined as a rational and conservative organization, in reality, governmental systems could be corrupted by people with personal egoistic goals, willing to take risks, privatize profits, and socialize losses. A government could be completely immersed in corruption, perhaps called a mafia-state (Naím, 2012). The main problem with such a corrupted organization is that its main goal is self-preservation and profit in the near-term mode, which lower quality of strategic decisions. One example is how Cambridge Analytica was hired by Russian oligarchs to manipulate elections in US and Britain, but these oligarchs themselves acted based on their local interests (Cottrell, 2018).

# Conclusion. Riding the wave of the AI revolution to a safer world

Any AI safety solution should be implementable, that is, not contradict the general tendency of world development. We do not have 100 years to sit in a shrine and meditate on a provable form of AI safety (Yampolskiy, 2016): we need to take advantage of existing tendencies in AI development.

The current tendencies are that Narrow AI is advancing while AGI is lagging. This creates the possibility of a Narrow AI-based strategic advantage, where Narrow AI is used to empower a group of people that also has access to nation-state scale resources. Such an advantage will have a small window of opportunity, because there is fierce competition in AI research and AGI is coming. The group with must make a decision: will it use this advantage for world domination, which carries the risk of starting a world war, or will it wait and see how the situation will develop? Regardless of the risks, this Narrow AI-based approach could be our only chance to stop the later creation of a hostile non-aligned superintelligence.

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.

Arbital. (2017). Advanced agent. Arbitral. Retrieved from https://arbital.com/p/advanced\_agent/

Archer, J. (2018, May 31). Google draws up guidelines for its military AI following employee fury. *The Telegraph*. Retrieved from https://www.telegraph.co.uk/technology/2018/05/31/google-draws-guidelines-military-ai-following-employee-fury/

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI and Society*, *31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y

Bostrom, N. (2002). Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology, Vol. 9, No. 1 (2002).*

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

Bostrom, N. (2017). Strategic Implications of Openness in AI Development. *Global Policy*, *8*(2), 135–148.

Brenton, L. (2018). Will Artificial Intelligence (AI) Stop Hacker Attacks? *Stay Safe Online*. Retrieved from https://staysafeonline.org/blog/will-artificial-intelligence-ai-stop-hacker-attacks/

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

Cottrell, R. (2018, March 27). Why the Cambridge Analytica scandal could be much more serious than you think: *The London Economic*. Retrieved from https://www.thelondoneconomic.com/opinion/why-the-cambridge-analytica-scandal-could-be-much-more-serious-than-you-think/27/03/

De Spiegeleire, S., Maas, M., & Sweijs, T. (2017). Artificial intelligence and the future of defence. The Hague Centre for Strategic Studies. Retrieved from http://www.hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf

Ding, J. (2018). Deciphering China’s AI Dream.

Faggella, D. (2013, July 28). Sentient World Simulation and NSA Surveillance - Exploiting Privacy to Predict the Future? -. *TechEmergence*. Retrieved from https://www.techemergence.com/nsa-surveillance-and-sentient-world-simulation-exploiting-privacy-to-predict-the-future/

Goertzel, B. (2012). Should Humanity Build a Global AI Nanny to Delay the Singularity Until It’s Better Understood? *Journal of Consciousness Studies, 19, No. 1–2, 2012, Pp. 96–111*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.3966&rep=rep1&type=pdf

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

Jena, M. (2017, September 11). OMG! CIA Has 137 Secret Projects Going In Artificial Intelligence. Retrieved April 10, 2018, from https://techviral.net/cia-secret-artificial-intelligence-projects/

Kerber, L. L., & Hardesty, V. (1996). *Stalin’s Aviation Gulag: A Memoir of Andrei Tupolev and the Purge Era*. Smithsonian Institution Press Washington, DC.

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/

Kushner, D. (2013). The real story of stuxnet. *IEEE Spectr. 50, 48 – 53*.

Lem, S. (1959). *The investigation*. Przekrój, Poland.

Love, D. (2014). Mathematicians At The NSA - Business Insider. Retrieved from https://www.businessinsider.com/mathematicians-at-the-nsa-2014-6

Maxwell, J. (2017, December 31). Friendly AI through Ontology Autogeneration. Retrieved March 10, 2018, from https://medium.com/@pwgen/friendly-ai-through-ontology-autogeneration-5d375bf85922

Midler, N. (2018). What is ‘Stanford Analytica’ anyway? – The Stanford Daily. *The Standford Daily*. Retrieved from https://www.stanforddaily.com/2018/04/10/what-is-stanford-analytica-anyway/

Millett, P., & Snyder-Beattie, A. (2017). Human Agency and Global Catastrophic Biorisks. *Health Security*, *15*(4), 335–336.

Muehlhauser, L., & Salamon, A. (2012). Intelligence Explosion: Evidence and Import. *Eden, Amnon; Søraker, Johnny; Moor, James H. The Singularity Hypothesis: A Scientific and Philosophical Assessment. Berlin: Springer.*

Murphy, M. (2018, April 9). Chinese facial recognition company becomes world’s most valuable AI start-up. *The Telegraph*. Retrieved from https://www.telegraph.co.uk/technology/2018/04/09/chinese-facial-recognition-company-becomes-worlds-valuable-ai/

Naím, M. (2012). Mafia states: Organized crime takes office. *Foreign Aff.*, *91*, 100.

Oberhaus, D. (2017). Watch ‘Slaughterbots,’ A Warning About the Future of Killer Bots. Retrieved December 17, 2017, from https://motherboard.vice.com/en\_us/article/9kqmy5/slaughterbots-autonomous-weapons-future-of-life

Ought. (2018). Factored Cognition (May 2018) | Ought. Retrieved July 19, 2018, from https://ought.org/presentations/factored-cognition-2018-05

Perez, C. E. (2017, September 10). The West in Unaware of The Deep Learning Sputnik Moment. Retrieved April 6, 2018, from https://medium.com/intuitionmachine/the-deep-learning-sputnik-moment-3e5e7c41c5dd

Preskill, J. (2012). Quantum computing and the entanglement frontier. *ArXiv:1203.5813 [Cond-Mat, Physics:Quant-Ph]*. Retrieved from http://arxiv.org/abs/1203.5813

Rosenbach, M. (2013). Prism Leak: Inside the Controversial US Data Surveillance Program. *SPIEGEL ONLINE*. Retrieved from http://www.spiegel.de/international/world/prism-leak-inside-the-controversial-us-data-surveillance-program-a-904761.html

Shaw, T. (2018, March 21). The New Military-Industrial Complex of Big Data Psy-Ops. Retrieved April 10, 2018, from https://www.nybooks.com/daily/2018/03/21/the-digital-military-industrial-complex/

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv:1712.01815 [Cs]*. Retrieved from http://arxiv.org/abs/1712.01815

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. (2018). Disjunctive scenarios of catastrophic AI risk. *Artificial Intelligence Safety And Security, (Roman Yampolskiy, Ed.), CRC Press*. Retrieved from http://kajsotala.fi/assets/2017/11/Disjunctive-scenarios.pdf

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

Statista. (2018). Number of Google employees 2017. Retrieved July 25, 2018, from https://www.statista.com/statistics/273744/number-of-full-time-google-employees/

Templeton, G. (2017). Elon Musk’s NeuraLink Is Not a Neural Lace Company. Retrieved February 14, 2018, from https://www.inverse.com/article/30600-elon-musk-neuralink-neural-lace-neural-dust-electrode

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The Art and Science of Prediction* (Reprint edition). Broadway Books.

Timerbaev, R. (2003). History of the international control of the nuclear energy. (К истории планов международного контроля над атомной энергией). *История Советского Атомного Проекта (40-е — 50-е Годы): Междунар. Симп.; Дубна, 1996. Труды. Т. 3. — 2003*. Retrieved from http://elib.biblioatom.ru/text/istoriya-sovetskogo-atomnogo-proekta\_t3\_2003/go,214/

Turchin, A. (2017). *Human upload as AI Nanny*.

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

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

Turchin, A., & Denkenberger, D. (2018a). Could slaughterbots wipe out humanity? Assessment of the global catastrophic risk posed by autonomous weapons. *Under Review in Journal of Military Ethics*.

Turchin, A., & Denkenberger, D. (2018b). Military AI as convergent goal of the self-improving AI. *Artificial Intelligence Safety And Security, (Roman Yampolskiy, Ed.), CRC Press*.

Turchin, A., Green, B., & Denkenberger, D. (2017). Multiple Simultaneous Pandemics as Most Dangerous Global Catastrophic Risk Connected with Bioweapons and Synthetic Biology. *Under Review in Health Security*.

Welchman, G. (1982). *The hut six story: breaking the enigma codes*. McGraw-Hill Companies.

Williams, B. (2017). Spy chiefs set sights on AI and cyber -. *FCW*. Retrieved from https://fcw.com/articles/2017/09/07/intel-insa-ai-tech-chiefs-insa.aspx

Williams, G. (2018, April 16). Why China will win the global race for complete AI dominance. *Wired UK*. Retrieved from https://www.wired.co.uk/article/why-china-will-win-the-global-battle-for-ai-dominance

Winston, A. (2018, February 27). Palantir has secretly been using New Orleans to test its predictive policing technology. *The Verge*. Retrieved from https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd

Yampolskiy, R. (2016). Verifier Theory and Unverifiability. Retrieved from https://arxiv.org/abs/1609.00331

Yampolsky, R., & Fox, J. (2013). Safety engineering for artificial general intelligence. *Topoi*, *32*, 217–226.

Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks*. (M. M. Cirkovic & N. Bostrom, Eds.). Oxford University Press: Oxford, UK.

Zetter, K. (2015). So, the NSA Has an Actual Skynet Program. *WIRED*. Retrieved from https://www.wired.com/2015/05/nsa-actual-skynet-program/

Митин, В. (2014). Нейронет (NeuroWeb) станет следующим поколением Интернета. *PC Week. Идеи и Практики Автоматизации*, *17*.