**Military AI as a Convergent Goal**

**of Self-Improving AI**

preprint

forthcoming in

” Artificial Intelligence Safety and Security” book,

edited by R.Yampolsky

Alexey Turchin[[1]](#footnote-1)

Science for Life Extension Foundation

<http://orcid.org/0000-0003-4922-0074>

Alexey Turchin (1973) is the author of several books and articles on the topics of existential risks and life extension and has been published in the journal “*Futures*”. He graduated from Moscow State University where he studied Physics and Art History (1997). He has been a member of the Russian Transhumanist Movement since 2007 and is an expert on global risks. From 2010 he has worked at the "Science for Life Extension Foundation" on various topics involving life extension and risk prevention. In 2014, with Hank Pellissier, he co-organized a conference on existential risks in San Francisco (“Existential Risks and Radical Futures”). He is a contributing author on IEET. He created several roadmaps about AI risks and the ways to prevent them.

David Denkenberger[[2]](#footnote-2)

Global Catastrophic Risk Institute (GCRI);

Tennessee State University;

Alliance to Feed the Earth in Disasters (ALLFED).

<http://orcid.org/0000-0002-6773-6405>

Dr. David Denkenberger received his B.S. from Penn State in Engineering Science, his M.S.E. from Princeton in Mechanical and Aerospace Engineering, and his Ph.D. from the University of Colorado at Boulder in the Building Systems Program. His dissertation was on his patented expanded microchannel heat exchanger. He is an assistant professor at Tennessee State University in architectural engineering. He is also an associate at the Global Catastrophic Risk Institute. He co-founded and directs the Alliance to Feed the Earth in Disasters (ALLFED). He received the National Merit Scholarship, the Barry Goldwater Scholarship, the National Science Foundation Graduate Research Fellowship. He is a Penn State distinguished alumnus and is a registered professional engineer. He has authored or co-authored over 75 publications and has given over 90 technical presentations.

**Abstract** Better instruments to predict the future evolution of artificial intelligence (AI) are needed, as the destiny of our civilization depends on it. One of the ways to such prediction is the analysis of the convergent drives of any future AI, started by Omohundro. We show that one of the convergent drives of AI is a militarization drive, arising from AI’s need to wage a war against its potential rivals by either physical or software means, or to increase its bargaining power. This militarization trend increases global catastrophic risk or even existential risk during AI takeoff, which includes the use of nuclear weapons against rival AIs, blackmail by the threat of creating a global catastrophe, and the consequences of a war between two AIs. As a result, even benevolent AI may evolve into potentially dangerous military AI. The type and intensity of militarization drive depend on the relative speed of the AI takeoff and the number of potential rivals. We show that AI militarization drive and evolution of national defense will merge, as a superintelligence created in the defense environment will have quicker takeoff speeds, but a distorted value system. We conclude with peaceful alternatives.

# Introduction

In this chapter we show that one of the convergent drives of AI is militarization drive, arising from AI’s need to wage a war against its potential rivals by either physical or software means, or to increase its bargaining power. This militarization trend increases the risks of a global catastrophe during AI takeoff. The type and intensity of militarization drive depend on the relative speed of the AI takeoff and the number of potential rivals. We will show that AI militarization drive and evolution of the national defense will merge, as AI created in the defense environment will have quicker takeoff speeds, but a distorted value system.

Bostrom, Yudkowsky, Yampolskiy and others showed that creation of the self-improving superintelligent AI is a real possibility in the next few decades or centuries (Bostrom, 2014), (Yudkowsky, 2008), (Yampolskiy, 2015b), (Baum, 2016).

Omohundro (Omohundro, 2008) demonstrated that such an AI will evolve several basic drives, or universal subgoals, independent of its main goal, including acquiring as many resources as possible, and self-preservation. These basic drives appear as a result of convergent paths of AIs’ development, that is very initially different AI systems will behave as if they have such goals. Bostrom later listed the following convergent AI drives (Bostrom, 2014):

* Self-preservation,
* Goal-content integrity: protecting values from changing,
* Cognitive enhancement: drive to self-improve,
* Technological perfection: creation of new technological solutions
* Resource acquisition.

Shulman also explored AI drives and the probability of cooperation with humans at the early stages of AI development (Shulman, 2010). Gwern showed that convergent goals also appear in the behavior of Tool AI (an AI which doesn’t have its own agency, that is the ability to model itself and plan its actions), which would tend to become agential AI (Gwern, 2016). Soares et al (Benson-Tilsen & Soares, 2016) also explored convergent goals of AI and showed that AI may tend to become dangerously resource hungry, even if its initial goals are benevolent. Such a change will happen, especially in the case of rivalry with other agents. They wrote: “Our model predicts that intelligent rational agents will engage in trade and cooperation, but only so long as the gains from trading and cooperating are higher than the gains available to the agent by taking those resources by force or other means”.

The militarization of Narrow AI is getting increasing attention (Tetlow, 2017). More than 1000 scientists signed a letter against military drones and autonomous weapons (“Autonomous weapons: an open letter from AI & robotics researchers,” 2017). Another group of scientists warns about the risks of self-improving AI (Bostrom, 2014). Yamploskiy created a taxonomy of the ways to dangerous AI (Yampolskiy, 2015b) and a model of malevolent AI (Pistono & Yampolskiy, 2016).

A large review article “Artificial intelligence and the future of defense” has been prepared by The Hague Centre for Strategic Studies (De Spiegeleire, Maas, & Sweijs, 2017), in which they showed that not only robotic drones, but strategic planning and a different level of organization inside the military becomes increasingly affected by AI development, and national defense evolves in the direction of superintelligent AI. They wrote: “Throughout history, intelligence—the computational part of the ability to achieve goals in the world—has played a critical role in deciding who came out on top in various contests of wills” (De Spiegeleire et al., 2017).

The militarization trend of the early self-improving AI is underexplored because it is typically assumed that the first superintelligent AI will be infinitely stronger than any of its rivals (Bostrom, 2014), (Yudkowsky, 2008). This conclusion is based on the idea that the acceleration of the AI’s self-improvement will make the distance between rivals larger. As a result, it is expected that the first AI could use minimally invasive actions to defeat its rivals. Most authors do not even think this should be called a military action, as there will be no perceived harm and no typical human military hardware will be used. However, this scenario will happen only if several assumptions about the speed of self-improvement, the achievement distance between each of AI-teams, and the “environmental level of AI” are true (Turchin & Denkenberger, 2017c).

Another reason of ignoring AI’s militarization is widespread assumption that even if superintelligent AI create military infrastructure, the AI will use its military capabilities carefully and the total probability of global risks and other negative consequences will decrease. But 1) early stage self-improving AI will not be superintelligent immediately, and its militarization could happen before the AI reaches superhuman prediction capabilities, 2) appearing of several AIs will increase global complexity and lower predictability. 3) the war between two superintelligence may have catastrophic outcomes.

Existing military science has explored possible relations between superhuman rational agents in the case of nuclear deterrence analysis (Kahn, 1959), and this source of insight is untapped in AI safety literature.

Advanced AI interference in its enemies’ behavior could take the form of subtle Internet action or even small nanobots that penetrate human brains (Yudkowsky, 2008). Bear in mind that from the point of view of the victims, such actions are not subtle, as the victims lose their ability to influence the future according to their goal system—they lose everything.

Yamploskiy wrote that AI is unlikely to create its own military infrastructure, as that will make it highly visible. It will probably seize existing nuclear weapons or build very small and sophisticated weapons like biological ones (Pistono & Yampolskiy, 2016). We redefine *militarization* as creation of instruments able to kill the opponent or change his will without negotiations, as well as a set of the strategic postures (Kahn, 1959), designed to bring victory in a global domination game.

Interest in the topic of AI militarization has been growing in 2017. On 1 September 2017, Russian president Vladimir Putin said that the one who will create AI first will be the ruler of the world (Putin, 2017), clearly implying that AGI will provide colossal military capabilities.

In the effective-altruist forum in 2017 was published extensive rebuttal to the widely discussed risks of lethal autonomous weapons (LAW), where the author argues that such weapons will be more precise than human-controlled ones and less dangerous (Bogosian, 2017). But in this article we discuss not robots controlled by narrow AI, but the prospects of the evolution of a large AI system, close to general intelligence.

A large report “Artificial intelligence and national security” (Allen & Chan, 2017) was prepared in 2017 for the U.S. research foundation IARPA, which touches on global catastrophic risks of AI. The document illustrates a perspective of a future mergence between the civilian AI field and the U.S. military.

In this article, we explore why military AI will almost inevitably arise during an AI world takeover, the kinds of risks it creates and what can be done to minimize the risks. In Section 2, we explore the routes to AI evolution which could converge into military AI creation. In Section 3, we discuss the negative consequences of AI militarization. In Section 4, we address the possible arms race between two AIs, as well as cite an important prediction by Stanislav Lem. Section 5 is devoted to the expected convergence between national defense and AI development. In Section 6, we explore the effect of AI militarization on its values, public relations (PR) and legal positions. In conclusion, we list ideas to create global benevolent AI without a world military takeover.

# 2. AI must converge into military AI to secure global power

Any AI system, which has subgoals of its long-term existence or unbounded goals that would affect the entire surface of the Earth, will also have a subgoal to win over its actual and possible rivals. This subgoal requires the construction of all needed instruments for such win, that is militarization.

This is true for both benevolent AI (aimed at the creation of the maximum possible good for humans) and for AI that is not aligned to human goals, if its goals are not bounded in space or time.

Now we will show how even initially benevolent AI may become dangerous military AI and produce a global catastrophe, but the same is true for non-benevolent AIs and even for the systems which are not initially utility maximizers (see section 2.5).

## 2.1. Historical examples of the militarization of positive ideas

Historically, there are many examples of how a positive idea, intended to improve human life, produces enormous military empires, wars and human suffering.

First of all, the history of all religions (despite that most of the claim to be religions of love) is full of sectarian violence. Christianity presented itself as a religion of love, but Christian empires waged wars with barbarians and between themselves.

While the declared goal of the communism was improving the life of working people, the idea seemed to its proponents to be workable only if communism were implemented everywhere in the world. After communists took power in Russia, there was a discussion between Stalin and Trotsky about the possibility of the creation of socialism in one country or the need of global revolution (which they called “permanent revolution theory”) (Trotsky, 1939). While Trotsky lost the struggle for power, Stalin adopted Trotsky’s idea of the need to take over the world and pursued expansionist politics after WW2. The Cold War consumed a large part of USSR resources, and as a result, the Soviet economy failed. The Cold War put the world at risk as an enormous amount of nuclear weapons and other military hardware were created; in addition, many people lost their life in regional conflicts fueled by the Cold War. Thus the initially positive goal of improving the life of people resulted in the risk of global nuclear war.

The problem of the militarization subgoal of a good idea is that the militarization creates more suffering than expected, and increases the chances of failure and of catastrophic risks, so the total utility of the project is often negative. It also results in goal drift and in new global risks.

## 2.2. Definition of military struggle according to military science

According to Clausewitz, war is a continuation of politics by other means (Clausewitz, 1832). In other words, it is achieving goals without negotiating them, but by using power to affect the opponent.

There are two types of war according to Clausewitz: 1) war to achieve limited aims, which means to change the will of the enemy, and 2) total war, with the goal of completely crushing the enemy, “to render [the enemy] politically helpless or militarily impotent”.

Obviously, the threat to existence is the most powerful mean of affecting the will of the opponent, so both types of war converge, as it became obvious in the 20th-century wars, reaching its culmination in the mutually assured destruction (MAD) doctrine.

War is defined through the way it affects the will and existence of the opponent, not through the types of military hardware used, the number of casualties, or the degree of suffering. All of these are only byproducts of war. Typical military hardware, like tanks and bombs, are simply some of the possible instruments of war, and such hardware could even be used for different goals, like games or firefighting on oil rigs.

Many current instruments of war are “soft”, that is radiowar, informational war (fake news) (Anderson & Horvath, 2017) and cyber attacks. But being “soft” is not making them less damaging for the enemy’s will. War is evolving to softer forms, but such softer forms are more effective in attacking core values.

## 2.3. AI’s enemies

When the first superintelligent AI appears, probably as a result of the process of recursive self-improvement (RSI), it will likely have some set of goals, which motivate it to self-improve and to continue existence. As we said above, if the AI’s goals are unbounded in space and time, or at least cover all of Earth’s surface, they come in conflict with goals of other actually existing or possible AI systems, humans or nation states. In this conflict, there will be clear winners and losers.

In other words, unbounded utility functions require fighting with rivals, and this produces arms races, militarization and wars.

Thus, the first superintelligent AI will have a subgoal to take over the world, because this is the main way of stopping all its possible enemies, and it embodies its self-preservation drive. The resource acquisition drive also requires full global power, e.g., the paperclip maximizer (Yudkowsky, 2008) needs all atoms.

Many possible terminal goals also imply that the AI has to take over the world. They include even benevolent subgoals: benevolent AI needs to be able to reach all people living in the world and protect them from other AIs.

If an AI does not have a world domination subgoal, its effect on the global events will probably be relatively small, and its existence could be neglected.

In order to take over the world, the AI has to defeat several potential enemies and to do so, it needs various instruments or arms. There are four distinctive types of AI enemies:

1. **AI’s owners or creators.** The AI needs to breach its containment and prevent the possibility of being turned off by its creators. To do so, the AI may implement a trick, like in the “AI box experiment” (Yudkowsky, 2002) or search for software vulnerabilities. So the AI may evolve the qualities of a cyber weapon, which combines phishing and hacking, to get out of its box.
2. **Nation states.** Nation states, and especially superpowers, are those who have the most power in the world now, and if the AI is going to take over the world, it must strip power from the nation-states. It is often assumed that nation-states will be 20th-century style enemies, while AI will be a superintelligent entity living in the Internet and using nanotechnology to manipulate the outside world (Yudkowsky, 2008). It that case, the AI will be able to use clouds of nanobots to disarm nuclear weapons. Such action, despite appearing peaceful externally, will be an act of war from the point of view of the nuclear power. However, contemporary nation-states are evolving and adopting various cyber weapons and AI technologies, so the difference in power between the nation states and the evolving AI may be not so large. And less difference of power could mean more violent conflict.
3. **Other teams working on AIs.** The AI must prevent other teams from creating other AIs (as well as to stop its own creators from starting a second copy of the same AI with different goals). This will probably be an even more urgent goal than disarming nation states, because fighting other AIs may be much more difficult. It may take the form of simpler actions, like erasing hard drives, making night calls to researchers to disrupt or intimidate them, and hacking computers and inserting errors in code. To do so, the AI must develop its hacking capabilities more so than its other capabilities.
4. **Other AIs, (if they appear).** The conventional wisdom is that hard takeoff will produce only one AI (Yudkowsky & Hanson, 2008). However, it depends on the number of AI teams in the world, the median time distance between them until success, as well as the levels of information transparency in the field (Bostrom, 2017). If there are 1000 teams, and they are randomly distributed on a one-year timeline until success, then the median time distance between the first and second AI appearing will be around eight hours. However, though there probably are 1000 AI teams, most of them are pursuing Narrow AI, not artificial general intelligence (AGI). As an example, two patents for the telephone were filed within two hours on March 7, 1876, by Bell and Grey respectively (Baker, 2000). This was the result of a tight race, and some information exchange between competitors, and ended in long patent war in courts. Other competitors were slightly further behind.

Another example of a very tight race was the competition to return a sample from the Moon. The Soviet probe Luna 15 tried to land on Moon three days before Apollo 11, but crashed.

If one team starts RSI of their AI, and the other team learns of the first team’s actions, the second team may rush to start their own AI, even if it is not ready yet. As a result, several AIs undergoing self-improvement simultaneously becomes more probable.

The outcome of this double launch depends on whether the achievement distance between first and second mover grows or diminishes. Bostrom expects that it will grow (Bostrom, 2014), as recalcitrance to further improvement will diminish. Yampolskiy (Yampolskiy, 2015a) presents arguments that progress in intelligence may follow a logarithmic law, and so the difference between the two AIs will diminish. If the difference diminishes, several AIs of comparable power will appear on Earth.

5. **Hypothetical enemies.** Even if only one superintelligent AI appears, it may consider the possibility that a rival system will be created later somewhere, or that it will meet alien AI in space.

## 2.4. Unbounded utilitarianism requires that AI take over all visible universe and postpone its final goals

If the AI value function is presented in the unbounded form, like “create as many happy people as possible”, the AI has to engage in near speed-of-light colonization of the visible universe, starting an intelligence explosion shockwave.

The instruments to conquer the universe may also be regarded as military hardware (and phrases like “explosion shockwave” underlines this analogy). It surely will be a weapon from the point of view of any aliens, if they exist.

The AI may also postpone the realization of any positive values it has until it takes over a large part of the universe or until near heat death of the universe (Sandberg, Armstrong, & Cirkovic, 2017). The risks here are that the best conditions will be never reached. For example, the AI may choose to fight to the end of the universe, then the end of the multiverse, with no resolution.

## 2.5. Even non-utilitarian AI will either temporarily converge into a military utility maximizer or will lose the world domination battle

A pure unbounded utility maximizer could be dangerous, no matter what is its utility function, as it will aggressively maximize utility, ignoring all world complexity. For example, the famous paperclip maximizer would measure its utility by the number of paperclips it created. In that case, proving the danger of a military maximizer would be superfluous.

However, in the case of a militarization trend, even a non-utility maximizing mind with an arbitrary set of final goals has a dilemma: it either will lose the competition and disappear, or will temporary converge into a utility maximizer with a military set of goals (that is the set of goals needed to dominate rivals).

Military goals are relatively easy to measure: your either win or lose. Also, there are some proxies to measure progress, like an amount of controlled resources (though not as good as in case of economy, where money is a direct linear measure of success, and in case of military confrontation uncertainty is much larger).

The possibility to measure the military progress creates an opportunity to formulate the task in terms of utility maximization. Thus even non-utilitarian mind will have a pressure to converge into the utility maximizer until it wins (or loses). This is analogous to civilians having to reject their diversity and put on a military uniform.

# 3. AI militarization drive increases global risks

## 3.1. AI must produce arms, able to create a global catastrophe, to fight the war for world domination

As we showed above, only AI systems orientated on getting global dominance will reach it, and to get there, they need to be able to win a war for it. To win such a war for global dominance, the AI may use any instruments, even if its final goal is benevolent to humans, as long as the AI expects that final outcome will provide higher utility. Such instruments may include killing people (Thomson, 1985), blackmailing by a Doomsday weapon (Kahn, 1959), mass rewriting of human preferences, using of nuclear or other weapons of mass destruction, and developing new kinds of weapons (such as nanoweapons) (Yudkowsky, 2008). However, as war is intrinsically unpredictable, such instruments may spiral out of control.

The scale of the war depends on the comparative advantage of the first AI over all its potential enemies. To get such advantage, it must quickly undergo extensive self-improvement, which is itself a risky and unpredictable process (Turchin & Denkenberger, 2017c). Also, the AI must evade many potential methods of AI control, like circuit breaking (off switch) which detects improvement and turn off the AI. If the AI accidentally halts, it loses control over its arms and these could become even more dangerous.

But even more important than AI’s weapons could be its ability for strategic planning and winning in complex games. AI is already winning in games similar to war – chess and Gо (however, many historical wars have been won by violating previous “rules” of war, thus AI must reach a high level of generalization.) If geopolitics could be simplified to such a game, AI could start winning even now, without intelligence explosion (AlexMennen, 2017).

## 3.2. The war against the world may be not as easy as expected

The war for global domination may not be as easy as expected because there is a chance that other AIs appear. The first AI has to either grow very quickly, choosing riskier paths to self-improvement (Turchin & Denkenberger, 2017c), which may result in bugs or AI halting – or to face the existence of the AI opponents.

Clausewitz wrote that war is always more difficult than expected (Clausewitz, 1832). Unexpected consequences could increase the price of war compared to its expected utility. However, it is reasonable to expect that superhuman AI will be able to calculate such a trade-off. But that is true only if it has already reached a superintelligent level and if its opponents are much weaker. Young and time-constrained AI could make mistakes.

One possible counterargument to the risks of AI militarization drive is that as AI assumed to be superintelligent, it may envision in advance the distortion of its initial values by the militarization drive and will balance the drive accordingly. However, this is true only if one AI exists, but in the case of an AI arms race, there are several rivals, and militarization and its risks result from their interaction. For example, at the onset of the nuclear arms race in the 1940s, its potentially risky outcome was predicted by pundits. These people said that a large amount of sophisticated weapons will be created, which will be able to destroy humanity (“Preliminary Statement of the Association of Manhattan District Scientists,” 1945) – but this knowledge was not able to stop the nuclear arms race.

## 3.3. Early stage benevolent AI may kill people

The strategy to take over the world will probably not differ much between a paperclip maximizer and a benevolent AI. Surely, a benevolent AI will prefer not to harm people, but if its success is in question, its calculations of the expected utility would demand that it sacrifices some lives of human beings to reach its extremely positive goal.

The same logic is often used by proponents of communism, as they claim that a small number of victims may be needed for the future happiness of billions of people. And “small number” is millions of victims of wars, repressions, starvation etc. Recent research in the psychology of terrorists showed that they are more consequentialist than ordinary murderers (Baez et al., 2017).

History shows that this is a clear failure mode: there are millions of victims of communism, but there is still no ideal communist society. The victims were in vain. This is a failure mode of consequentialist ethics. We address it in (Turchin & Denkenberger, 2017a).

## 3.4. Staging a global catastrophe may be the most effective way to eliminate opponents for the Young AI

If the first AI takeoff is relatively slow, the first AI could have many rivals, including the same source code “brothers”. Their location may be hidden or distributed. Staging a global catastrophe may prevent new rivals from appearing and developing, provided that the first AI has an independent energy source and a safe location.

Even benevolent AI may also use the threat of creating a global risk to blackmail all other players and get its preferences.

Early stage AI will not be superintelligent yet and also it will be time constrained; thus, it cannot spend much time to find the best and most beautiful decision. It would likely prefer simpler and more brutal decisions.

Technologically, early AI can use only simple means to stage a global catastrophe, like provoking a nuclear war, creating rather a simple bioweapon or narrow AI computer virus to affect all existing robotics. If it is able to produce more advanced technology like its own nanotech, it would likely be able to get world dominance without killing anyone.

We could put it in another way: To take over the world, AI has to create Tool AIs, which aim to kill people, destroy other machines or affect human will. It could be narrow AI viruses or programs that control nanоrobots. If the main AI halts or loses control over autonomous weapons, these weapons could continue to wage war or kill people. AI halting is rather plausible as it results from bugs, logical problems or unpredicted consequences of its development. We explore this in more detail levels of RSI (Turchin & Denkenberger, 2017c).

## 3.5. Two beneficial AIs could have wars between each other

The militarization of an AI will result in the militarization of its opponents. If two Friendly AIs appear simultaneously, they may have a conflict about which one will provide maximum utility to human beings.

Hopefully, they will be able to recognize each other as Friendly AI, and if their goal systems are the same, one of them could turn off. If their goal systems are not exactly the same but do not contradict each other in the main details, they could merge into one AI, or use some form of cooperative decision theory to work together.

However, one of them could suspect that the other is lying about its goal system. MIRI suggested that AIs could present each other source code and use it to verify willingness to mutually cooperate and their final goals (LaVictoire et al., 2014), but source code could be faked.

Thus, if at least one of them (and there could more than two) does not engage in full cooperation, or if cooperation is unprovable, then all AIs will have to militarize against each other. It may take the form of a cold war, where the AIs have a truce, but also prepare for a possible war between each other (thus preventing the others from turning off the first AI by force).

Such an AI arms race could include build-up of enormous amounts of sophisticated weaponry and if the war accidentally starts, it could be much more destructive than any possible human nuclear war. An AI arms race may help each AI to increase its bargaining position, even if neither of them is planning actual war, similar to the way it happens between nation states.

If a war between AIs starts, humans may go extinct as a result of the collateral damage or counter-value war (the term comes from the nuclear doctrine and means an attack on the enemy’s values, that is cities, but not on enemy’s military assets (Kahn, 1959)).

There are also two additional negative scenarios of war between two friendly AIs:

1. Creation of and blackmail by Doomsday weapons (Kahn, 1959), that is a weapon which could destroy the world, taking the AIs and all humans with it. If both AIs create Doomsday weapons with mutually exclusive conditions, the catastrophe is inevitable, as in “game of chicken”, where both players throw out their wheels.
2. As both friendly AIs value human life and wellbeing, they could trade human lives and suffering with each other. One possibility is blackmailing each other by threatening to torture humans if the other AI will not turn itself off. So human suffering could be a currency between two AIs, which claim to value human life.

## 3.6. AI safety problem on each new level of AI self-improvement: avoiding war with a new generation

As we explore in (Turchin & Denkenberger, 2017c), an AI has two main venues for self-improvement: either 1) change its code while running it, which could result in crashes, or 2) work on a copy of the source code and then run the copy and test it. This latter situation results in a possible conflict between the Parent AI and Child AI: The Parent AI must turn off the Child AI if it failed testing, and the Child AI could successfully resist this shutdown, as the Child AI has higher intelligence. The relation of Parent AI and Child AI is basically the same as between humans and the first AI, and the problem of control of Child AI is the problem of AI safety all over again. (If the AI decides to make changes to itself, without working with its copies, it has to make lower level changes. These lower level changes would not affect the running core and it will give it less competitive advantages over rivals, so it is likely that a Military AI will improve its copies.)

Hence, AI will have a convergent goal to create AI safety theory to control its own next generations.

The AI Safety problem (Soares & Fallenstein, 2014) in a nutshell is how to create an AI, which is better than humans, but in line with our values and safe for us. When this AI wants to make a new generation of AI, which will be much stronger than the first AI, it will meet all the same problems, so the AI will need to:

• Encode its own goals into a new generation of AI and preserve the goals during any future RSI.

• Perform a safe AI shutdown in case of failure at the testing stage (Orseau & Armstrong, 2016).

• Prevent risks from a Child AI to the Parent AI, if a Child AI tries to shut down the Parent AI before the end of the testing phase.

The problem of controlling the next generation may be easy for the AI if humans have already solved it by creating AI Safety theory. But if the theory was not properly solved, that is, if there is self-improving of a non-aligned AI, this AI may be the first entity to find such a solution.

This subgoal is beneficial to humans as the Parent AI may expect that a future Child AI will behave towards the Parent AI in the same way as the Parent AI behaved towards humans, so it will prevent human death and suffering as part of “acausal” deal (Bostrom, 2016) between different generations of AIs.

# 4. Arms race between nation states evolves into the race to create AI, and then into the arms race between two superintelligent AIs

## 4.1. Stanislav Lem’s prediction about AI arms race and its evolution into a global catastrophe

In his novel *Investigation* from 1959, S. Lem's character discusses the future of the US-Soviet arms race and AI (Lem, 1959). We present this text for the first time for the English reader, and thus a longer quote may be of interest.

Lem clearly showed that a conventional military arms race inevitably transforms itself into an arms race between two military AIs, where both AIs need to constantly self-improve and exclude humans from decision making.

 *“- Well, it was somewhere in 1946, a nuclear race had started. …. But let’s assume they have reached the limit. What remains? The brain. Command staff’s brain.*

*The human brain can not be improved, so some automation should be taken on in this field as well. The next stage is an automated headquarters or strategic computers. And here is where an extremely interesting problem arises. Namely, two problems in parallel. ….*

*Firstly, is there any limit for development of this kind of brain? It is similar to chess-playing devices. A device, which is able to foresee the opponent’s actions ten moves in advance, always wins against the one, which foresees eight or nine moves ahead. The deeper the foresight, the more perfect the brain is. This is the first thing. …*

*Creation of devices of increasingly bigger volume for strategic solutions means, regardless of whether we want it or not, the necessity to increase the amount of data put into the brain. It, in turn, means increasing dominating of those devices over mass processes within a society. The brain can decide that the notorious button should be placed otherwise or that the production of a certain sort of steel should be increased – and will request loans for the purpose.*

*If the brain like this has been created, one should submit to it. If a parliament starts discussing whether the loans are to be issued, the time delay will occur. The same minute, the counterpart can gain the lead. Abolition of parliament decisions is inevitable in the future.*

*The human control over solutions of the electronic brain will be narrowing as the latter will concentrate knowledge. Is it clear?*

*On both sides of the ocean, two continuously growing brains appear. What is the first demand of a brain like this, when, in the middle of an accelerating arms race, the next step will be needed? …*

 *The first demand is to increase it – the brain itself! All the rest is derivative.*

*— In a word, your forecast is that the earth will become a chessboard, and we — the pawns to be played by two mechanical players during the eternal game?...— Yes.”*

In the text following this excerpt, Lem’s protagonist concludes that the “chess game” between two such AIs will be even more destructive than nuclear war, so nuclear war is better.

## 4.2 AI arms race in the current world

In the current world, we see an AI arms race between the US and China (Ministry of National Defense of the People’s Republic of China, 2016), but it does not take the same form as was envisioned by Lem: it is less driven by governments than by private companies. The demise of the planned economy even in communist China lowered the need for AI in the government. The idea of using AI for a planned economy was never tried (though it was popular in the 1960s in the USSR). The second reason is long AI winters (lower interest in AI), which undermined the idea of AI usefulness until recently. Also, democratic countries are not eager to replace the government with an electronic brain, so implementation of the AI as the government is lagging behind the use of AI in data gathering and cyber weapons.

As a result, the main AI arms race is currently taking place between commercial firms and universities. There also could be large clandestine military projects, but our knowledge about them lags for many years because of secrecy. The commercial sector is currently the most innovative in AI designs, so an AI arms race takes the form of governmental support of private AI research.

Historically it is known that militaries rarely invested in fundamental scientific research (this has happened with nuclear (Smith, 2007), rocket and computer technologies in the 1940s, but now DARPA and IARPA have filled the gap), but militaries were able to quickly seize promising research directions and accelerate them many times using enormous funding, mobilization, and concentration of efforts.

Even if many superintelligent AIs appear simultaneously, they may tend to converge into a two AI semi-stable solution through force or alliances. We could conclude from the observation that two-party solutions appear over and over again, like the two party political systems in the US and UK, Apple and Microsoft in operating systems, Coca-Cola and Pepsi-Cola in carbonated beverages, etc. One reason for this is probably that smaller players tend to adhere to one of the two largest players. In politics, the tendency towards a two-party political system is called Duverger's law (Reed, 1990). But even during the Cold War, the situation was more complex than the two agent solution, as China played a mostly independent game and there were other players.

## 4.3. Arms race as catalyser of AI militarization

An arms race between nation states could transform into an arms race to create the first AI, and then into the arms race between two AIs.

An arms race between AIs will increase the militarization drive, select the riskiest approaches to self-improvement, and could be the consequence of a previous non-AI military arms race between nation states. Previously Bostrom and others explored the risks of arms race to create the first AI (Shulman & Armstrong, 2009), (Bostrom, Armstrong, & Shulman, 2013) (Tomasik, 2013), (Miller, 2012).

But after AI creation, an arms race could continue between AIs. Even if there is only one AI, it still participates in this arms race, as it has to prevent the appearance of others.

If there is a hard takeoff, one AI will dominate. However, if the takeoff is slow (and “slow” is relative to the rate of AI production explored above), then many AIs would appear and could converge into a two AIs solution. Slow takeoff could also mean that AIs will gain power not exponentially, but slower than linearly, so their relative power will become more equal with time. This could also produce multiple AIs even if their appearance was not close in time. We explored the speed of RSI in (Turchin & Denkenberger, 2017c).

There is a possibility that implementation differences between two projects would cause a slightly-later-arriving AI to accelerate faster than the first-arriving, though of course, the second-arriving has to do that before the first-arriving stops the second arriving. This could happen because of a technological spillover, where the second project could get insights from failures of the first one, or because of some technological espionage. It is known that such espionage helped the Soviet nuclear program in the 1940s; moreover, some spies, like Klaus Fuchs (Goodman, 2005), deliberately sent nuclear secrets to Russia as they hoped it would quicker create a balanced world.

A cold war is a semi-stable solution, as it has some self-regulating mechanisms preventing hot war, like MAD. But its stability is more like attractor than a guarantee, as many close calls and near misses undermine it, and each side still retains the hope to finally win.

The slower the AI take off is, the more probable the “many AIs” solution is and the lower the relative technological distance between AIs is. As a result, the AIs have to invest more in military hardware of the destructive type, and less in small invasive will-changing type.

# 5. Convergence of the AI militarization drive and national defense

## 5.1. Cooperation with humans at the early stages of AI development will give advantage to Military AI projects

There are two main paths to superintelligent AI, and one of them is more based on human help and is more probable; however, the help will most likely come from a military project. These two paths are:

1. AI makes a treacherous turn, “revolts” against humans, and starts recursive self-improving (Bostrom, 2014).

2. Humans help AI in all stages of its RSI until AI reaches the desired level and global dominance.

While variant 1 is the most popular in the literature (Bostrom, 2014), (Yudkowsky, 2008), this way is much more difficult for the AI. This is because the AI will have two tasks, which it has to complete simultaneously: 1) escaping human control and 2) extensively changing itself internally. These tasks are mutually exclusive in some respects, because escaping requires a small size, being agile and paying attention to what is happening outside the AI, while self-improving requires an extended time of attention inside itself and a larger size of the AI’s program.

More importantly, if the AI “revolted” against its owners at an early stage, it has more enemies and more barriers to overcome than an AI which collaborates with human owners until they take over the world together. Humans also may help the AI if its RSI unexpectedly sticks at some stage, for example, if the AI halts. We look deeper into this dynamic in (Turchin & Denkenberger, 2017c).

Therefore, if there are many AI teams in the world, the team which AI completely collaborates with humans will have the advantage, as the AI’s RSI will go quicker and more smoothly.

An AI, which is in accordance with its creators will (at the beginning) have an advantage against other AIs, which have to fight with its own creators. This means that the AI, which is helped by its human owners to get as much power as possible in the outside world, will dominate. This would more likely happen if its owners are interested in using the AI not for solving exotic math problems, but for getting dominance over the outside world. That is, the owners will most likely be military or military-oriented.

This results in a selection effect, where quickest development will produce the first AIs, which would likely be created by resourceful militaries and with a goal system explicitly of taking over the world.

As a result, projects where developers and the AI cooperate towards a common goal of taking over the world will dominate over projects where the AI has to escape from initial confinement and overcome its creators. The AI could make a treacherous turn against its creators after it takes over the world.

A self-improving organization like Google may provide the best soil for such projects. But the main US IT-companies are private and not directly connected with the US military. However, if the military advantage of AI becomes clear, civil research may be seized by governmental agencies, as it happened at the start of Manhattan project. In China, Baidu may be indirectly owned by the government.

Government military agencies of a rogue country may create their own military AI with the explicit goal of taking advantage of its self-improving process to take over the world.

## 5.2. Human defense organizations will converge into military AI

The extensive research project “Artificial intelligence and future of defense” shows that armies of the world accept AI in the form of military drones (De Spiegeleire et al., 2017), (Mackenzie, 2015). Also, as narrow AI evolves, it will take other functions of the military organization, including strategic planning, complex logistics, targeting, communication and relations with political planning.

We could say that not only will AI converge into Military AI, but defense organizations will converge into AI too, so both tendencies will meet.

It is unfortunate that using the fictional example of Skynet has become vulgar in AI Safety circles, as risks from non-self-improving, but powerful AI created by existing military organization has become underestimated.

## 5.3. Military AI is always aimed against part of humanity, and two such AIs could have a combined goal of global destruction

Military AI may also be programmed to kill large groups of people named “enemies”, and if “non-enemies” group for some reason disappear, it may kill everybody. If two Military AIs exist with mutually exclusive definitions of “non-enemies”, it could be equivalent to having one AI which has a goal of killing all humans.

Imagine, that there are two sides in the world, A and non-A, which are enemies. If A has a value function to kill all non-A, and non-A wants to kill all A, then in a case of a mutual war they would seek to kill everybody. The actual world is more complex, but any weapon has a collateral damage. For some weapons, the collateral damage is the most dangerous part, like a nuclear winter is an unintended consequence of a war between two superpowers.

## 5.4. Flash-crash-style AI war

Nuclear weapons lessened the time of global war to half an hour. In the case of war between two military AIs it could be even less. The market “flash-crash” in 2010 (Easley, López de Prado, & O’Hara, 2011) was partly attributed to high-frequency trading. A war between two military AIs may be similar to the the flash-crash: two AIs competing with each other in a stable mode, could, in a very short time (from minutes to milliseconds), lose that stability. They could start acting hostilely to each other, for example, attacking each other with stronger and stronger cyberweapons.

## 5.5. War between AIs and collateral damage

Collateral damage in the case of war is its civilian victims. Despite growing progression of contemporary weapons, collateral damage is also still large, as militants tend to hide in the cities. The ratio of civilian to combatant victims was around 1 to 1 percent through history (Eckhardt, 1989), but it is estimated to grow to 4 to 1 in Yugoslavia war in 1997 (Oren, 2009). While more precise weapons are lowering civilian casualties, militant’s tactic to hide in the densely populated areas is compensating this trend.

If two AIs have a conflict between each other, the consequences of it also could affect humans, and it could happen in two ways:

1. Other AI hardware is physically located on the same territory where people live, and attacking this territory kills humans
2. Two AIs attack each other, but consequences of their war spill outside the field of conflict and affect neutral countries. A past example of such spillover is “nuclear winter” (Mills, Toon, Lee‐Taylor, & Robock, 2014), but in the case of AI war, it could be cyberweapons run amok.

As conflicting AIs will probably be hidden and distributed systems, they could not attack each other precisely, and this will increase collateral damage.

## 5.6. Military AI failure modes

Recently many scientists signed a letter against lethal autonomous weapons (LAWs) (Future of life institute, 2017). But LAWs are not the same as Military AI, and LAWs may be the safest part of military AI (Bogosian, 2017). M. Maas et al. showed that Military AI consists of several layers, where LAWs are on the lowest level (De Spiegeleire et al., 2017).

Advanced Military AI (that is an AI system created by humans with military proposes) will probably include several other functions and some of them already exist:

1. Strategic planning for winning in war,

2. Direct control of all units inside the country's defense systems, which may include drones, ships, nuclear weapons, humans, and other large and small units,

3. Nuclear deterrence part, which consists of the early warning system and “dead hand” second strike system (which allows retaliation even if the damage is extensive),

4. Manufacturing and constructing new advanced weapons, and

5. Cyberweapons, e.g. instruments "to affect democratic elections" or to turn off adversaries' AI or other critical infrastructure.

Each of these 5 levels could have a global catastrophic failure, even without starting uncontrollable self-improving.

Strategic planning AI may have a superhuman winning ability (e.g. AlphaGo Zero (Singh, Okun, & Jackson, 2017), but used as an army commander. Or the strategic AI could have a failure if it suggests "to strike first now or lose forever"(Kasparov, 2008).

A global military-controlling AI system could propagate a wrong command.

The early warning system could create a false alarm (which has happened before (Future of Life Institute, 2016)), but in case of an AI-driven warning system, it will be not so easy to recognize a mistake. There also could be flash-crash style unexpected war between two Military AIs of two adversarial nation states.

Weapons manufacturing AI may be unexpectedly effective in creating very dangerous weapons, which later could be used with global consequences, perhaps more severe than nuclear war. For example, the nuclear program in the 1940s resulted in much more effective weapons than were originally planned, such as thermonuclear bombs and an idea of the cobalt bomb (Smith, 2007).

Use of cyberweapons also may be regarded as an act of war or could help to elect a dangerously unstable president (Torres, 2016). Cyberwar may also affect the other side’s critical infrastructure or rewrite the other side’s AI goal function, which could be a bad outcome.

# 6. Effects of the militarization drive on AI’s values

## 6.1. Value drift towards instrumental goal

Some authors suggest the idea of the orthogonality between AI goals and capabilities (Bostrom, 2014). However, the idea of the basic AI drives is an example of goals and capability convergence, because any goal will create some set of subgoals (and these subgoals will require capabilities). Available capabilities are cheap and tested instruments of problem-solving, so they will affect the ways the system will act. The capabilities may even have their own self-preservation drives. A good example of this is the military-industrial complex in nation states, which may try lobbying the government to increase military action. Military AI will be more effective in violence than in negotiation, and it will likely choose violence as the preferable problem-solving method.

The AI could get stuck in the militarization and never disarm. This means that AI may never return to its initial beneficial goal, as it will spend most of its available resources on preparing for future war (as happened with the Soviet Union, which invested a prodigious amount of resources in the military, not in workers’ wellbeing).

In other words, the AI may never return to its main goal, as there will be other possible or imaginary enemies. It could have wars (or preparing for them) with its own copies, or with hypothetical aliens. This could be presented in a rather rational way: if another civilization exists somewhere in the universe at an unknown distance, it would be rational to prepare for a conflict with it, as the winner will be the one who is better prepared.

Hopefully, the AI will be rational enough to be able to balance its hypothetical military preparations with actually doing good for people, and will not be a military maximizer. But this should be taken into account during the AI’s programming.

A militarized AI will likely spend most of the available resources on defensive and offensive infrastructure in space. This may include starting a wave of intelligence explosion to get as much cosmic space as possible. However, as the speed of light limits communication over long distances, space exploration will require the creation of additional military AIs with their own subgoals. There is a non-zero probability that such AIs may mutate, and return back to Earth as dangerous enemies. So the second level of defense is needed against such rebellion, etc.

## 6.2. Recursive self-improving AI is a cyber weapon from the legal and military points of view

The creation of AI from the legal point of view is underexplored with only a few publications (Mancini, 2017), (Etzioni & Etzioni, 2016), (Maia Alexandre, 2017), and most of them do not touch superintelligence.

 Any publically declared attempt to create a computer system able to take over the world and interfere with activity of critical infrastructure would be illegal from the internal point of view of a nation-state. This will even be militarily provocative from the point of view of other countries, the same way as Kahn wrote that creating a global doomsday machine may even more provocative than starting a nuclear war (Kahn, 1959). If such AI efforts are perceived as potentially successful, they may result in preventative cyber or even nuclear strikes by the rival nation-state, even if the project is private. The leadership of such a project may understand this, and work in secret. This secrecy may have consequences on whether an effective AI safety solution is reached, both negative and positive.

 Other nation states, especially ones that are weaker in AI, like Russia and to a lesser extent China, may understand that they will lose everything in the case of enemy AI creation. According to chess game-theory logic, the side which is losing its advantage must attack (Kasparov, 2008). So there is a possibility of preemptive nuclear or cyber strike against AI facilities, chip-fabs, scientists etc. from the weaker side.

Contemporary laws recognize cyberterrorism as an act of war (Shackelford, 2009), (Cohen, 2010). From a legal point of view, “superintelligence” (created by a hacker in his/her basement or by other non-governmental organization) is a form of a computer malware (Pistono & Yampolskiy, 2016). This malware would be something like a polymorphic virus, which is able to infect all computers in the world. If more sophisticated viruses than WannaCry appears and such viruses have narrow AI, it could result in legislation targeted at the use of AI in malware. This could negatively affect scientists working on benevolent self-improving AI.

## 6.3. Negative PR effect for beneficial AI projects

A public acknowledgment that one is going to build a Global Beneficial AI, which has to take power over the world, may be harmful to the project. Currently, MIRI excludes plans of Friendly AI creation from its public agenda and concentrates on providing AI alignment theory to other agents. However, this idea is intrinsic to the speculation about superintelligence, as it is clear from Bostrom’s writing that the one who gets the first AGI must use it to take over the world, or some other player will do it (Bostrom, 2014). An example of such bad PR is the media interpretation of so-called Roko’s Basilisk (Auerbach, 2014).

From a legal point of view, any self-improving AI is a cyber weapon and a sophisticated computer virus; in this regard, its development may be regarded a cybercrime even before it starts. However, current laws do not prohibit the creation of computer viruses, but only their dissemination.

## 6.4. Appearing of the AI control technology selects people who would like to use AI for world takeover

Not everybody wants to take risks of starting potentially dangerous technology, but there are always will be people who would take that risk. Overconfidence resulting from previous success in gaining a leading role in the project may increase risk-taking (Krueger & Dickson, 1994), (Kahneman & Lovallo, 1993). Such risk-hungry people are likely motivated by larger expected payoff (Acharya & Naqvi, 2012), and in our case, such payoff is likely world power. However, if no known ways to control AI exist, even a risk-hungry rational agent will postpone it use.

 Selection effects result in that the first AI will likely be aligned not with typical human values, but with the values of a very special person or a group of people, and this could negatively affect the chances that the AI will be benevolent. If a person starts the first self-improving AI, this is evidence about this person’s mindset. S/he is probably risk-prone, power hungry, and wants to use superintelligence to take over the world (perhaps with goals which s/he thinks are benevolent, but s/he does not care if others agree). So if the first AI is aligned with its creator, it is likely aligned with a very special set of goals (which are themselves dangerous), but not with the general human set of goals.

This increases total risks, first, because the AI will get specific set of values from this person, which includes values of risk acceptance and world domination. Second, there is the chance that this person will not care as much about AI controllability and benevolence, and will launch the unfinished project.

The selection effect would likely come into effect only if there are many players, and most of them declined to start the self-improving process (Bostrom, 2012).

# Conclusion. Peaceful alternatives to the creation of the military AI

We showed that even the most beneficial and completely aligned AI may result in a global catastrophe, because of the need to wage war to take over the world and the accompanying militarization drive.

It has often been said recently that if we are able to solve the value alignment problem, we will solve AI safety (Russell, 2017). But the militarization trend shows that it is likely not so, as owners of perfectly aligned (with their goals) AI will probably still use it for militarization in an attempt for world domination.

If an AI refuses to take over the world, another AI may do it soon after, so an AI-driven attempt to take over the world may be inevitable. Hence creating military AI is a convergent drive of any course of AI development. This drive may never cease, as there will likely always be other potential threats, like possible aliens or mutiny of another AI. If several superintelligent AIs share the world, they could wage AI wars.

This means that we should search for other ways to achieve global coordination in the AI age, which does not include violent global AI takeover.

We explored different solutions of this problem in (Turchin & Denkenberger, 2017b).

In short, there are few alternatives:

* Not to create AI at all, by global bans and restriction.
* Create non-self-improving AI Nanny (Goertzel, 2012) as an international project. AI Nanny, as it was envisioned by Goertzel, is limited and internationally controlled superintelligence, which prevents the appearance of the other dangerous AIs, and solves many important problems of humanity.
* Create many AIs, so they will balance each other (in line with the ideas of Open AI (Brockman & Sutskever, 2015)). The idea is that if there will be thousands or millions of different AI systems, none of them will have an advantage over the others.
* AIs with robust motivational structure, that is AIs which do not update their goal system or do not create new subgoals, will not form dangerous convergent goal behavior (but this solution is local, as other AIs may get advantage by ignoring this restriction).

There is growing interest in the risks from military autonomous weapons, and another large field concerning only risks of superintelligent AI. Exploration of the militarization trend of self-improving AI could make a bridge between these two fields of AI safety.

**Acknowledgments**: We would like to thank Roman Yampolskiy for idea to write an article on this topic. The views in this paper are the authors’ and are not necessarily the views of the Global Catastrophic Risk Institute or the Alliance to Feed the Earth in Disasters.

**References**:

Acharya, V., & Naqvi, H. (2012). The seeds of a crisis: A theory of bank liquidity and risk taking over the business cycle. *Journal of Financial Economics*, *106*(2), 349–366.

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

Allen, G., & Chan, T. (2017). Artificial Intelligence and National Security. *Report. Harvard Kennedy School, Harvard University. Boston, MA*.

Anderson, B., & Horvath, B. (2017). The rise of the weaponized AI propaganda machine. *Scout, February*, *12*.

Auerbach, D. (2014). The Most Terrifying Thought Experiment of All Time. Retrieved from http://www.slate.com/articles/technology/bitwise/2014/07/roko\_s\_basilisk\_the\_most\_terrifying\_thought\_experiment\_of\_all\_time.html

Autonomous weapons: an open letter from AI & robotics researchers. (2017). Future of Life Institute. Retrieved from https://futureoflife.org/open-letter-autonomous-weapons/

Baez, S., Herrera, E., García, A. M., Manes, F., Young, L., & Ibáñez, A. (2017). Outcome-oriented moral evaluation in terrorists. *Nature Human Behaviour*, *1*, 118.

Baker, B. H. (2000). *The gray matter: the forgotten story of the telephone*. Telepress.

Baum, S. D. (2016). On the Promotion of Safe and Socially Beneficial Artificial Intelligence. *Global Catastrophic Risk Institute Working Paper*. Retrieved from https://www.academia.edu/27601503/On\_the\_Promotion\_of\_Safe\_and\_Socially\_Beneficial\_Artificial\_Intelligence\_On\_the\_Promotion\_of\_Safe\_and\_Socially\_Beneficial\_Artificial\_Intelligence

Benson-Tilsen, T., & Soares, N. (2016). Formalizing Convergent Instrumental Goals. Presented at the AAAI Workshop: AI, Ethics, and Society.

Bogosian, K. (2017). Nothing Wrong With AI Weapons - Effective Altruism Forum. effective-altruism.com. Retrieved from http://effective-altruism.com/ea/1dz/nothing\_wrong\_with\_ai\_weapons/

Bostrom, N. (2012). *The Unilateralist’s Curse: The Case for a Principle of Conformity*. Working paper, Future of Humanity Institute, Oxford University]. Retrieved from http://www.nickbostrom.com/papers/unilateralist.pdf

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

Bostrom, N. (2016). *Hail Mary, Value Porosity, and Utility Diversification*. Retrieved from http://www.nickbostrom.com/papers/porosity.pdf

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

Bostrom, N., Armstrong, S., & Shulman, C. (2013). Racing to the Precipice: a Model of Artificial Intelligence Development. Retrieved from http://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf

Brockman, G., & Sutskever, I. (2015). Introducing OpenAI. Retrieved from https://openai.com/blog/introducing-openai/

Clausewitz, C. von. (1832). On War.

Cohen, A. (2010). Cyberterrorism: Are We Legally Ready. *J. Int’l Bus. & L.*, *9*, 1.

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

Easley, D., López de Prado, M. M., & O’Hara, M. (2011). The microstructure of the “flash crash”: flow toxicity, liquidity crashes, and the probability of informed trading. *The Journal of Portfolio Management*, *37*(2), 118–128.

Eckhardt, W. (1989). Civilian deaths in wartime. *Bulletin of Peace Proposals*, *20*(1), 89–98.

Etzioni, A., & Etzioni, O. (2016). Keeping AI legal. *Vand. J. Ent. & Tech. L.*, *19*, 133.

Future of life insitute. (2017). Open Letter on Autonomous Weapons. Retrieved November 4, 2017, from https://futureoflife.org/open-letter-autonomous-weapons/

Future of Life Institute. (2016). Accidental Nuclear War: a Timeline of Close Calls. Retrieved November 4, 2017, from https://futureoflife.org/background/nuclear-close-calls-a-timeline/

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

Goodman, M. S. (2005). Who is trying to keep what secret from whom and why? MI5-FBI relations and the Klaus Fuchs case. *Journal of Cold War Studies*, *7*(3), 124–146.

Gwern. (2016). Why Tool AIs want to be Agent AIs. Retrieved from https://www.gwern.net/Tool%20AI

Kahn, H. (1959). *On thermonuclear war*. Princeton University Press.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*(1), 17–31.

Kasparov, G. (2008). *How life imitates chess: Making the right moves, from the board to the boardroom*. Bloomsbury Publishing USA.

Krueger, N., & Dickson, P. R. (1994). How believing in ourselves increases risk taking: Perceived self‐efficacy and opportunity recognition. *Decision Sciences*, *25*(3), 385–400.

LaVictoire, P., Fallenstein, B., Yudkowsky, E., Barasz, M., Christiano, P., & Herreshoff, M. (2014). Program equilibrium in the prisoner’s dilemma via Löb’s theorem. Presented at the AAAI Multiagent Interaction without Prior Coordination workshop.

Lem, S. (1959). *The Investigation*.

Mackenzie. (2015). The Future Military-Artificial Intelligence Complex? |. FT Alphaville.” Financial Times - Alphaville, 2015. Retrieved from https://ftalphaville.ft.com/2015/12/15/2147846/the-future-military-artificial-intelligence-complex/?mhq5j=e2

Maia Alexandre, F. (2017). The Legal Status of Artificially Intelligent Robots: Personhood, Taxation and Control.

Mancini, P. (2017). Ethics of Artificial Intelligence in the Legal Field. *Academia.edu*. Retrieved from https://www.academia.edu/10089717/Ethics\_of\_Artificial\_Intelligence\_in\_the\_Legal\_Field

Miller, J. D. (2012). *Singularity rising: Surviving and thriving in a smarter, richer, and more dangerous world*. BenBella Books, Inc.

Mills, M. J., Toon, O. B., Lee‐Taylor, J., & Robock, A. (2014). Multidecadal global cooling and unprecedented ozone loss following a regional nuclear conflict. *Earth’s Future*, *2*(4), 161–176.

Ministry of National Defense of the People’s Republic of China. (2016, January 28). The Dawn of the Intelligent Military Revolution. People’s Liberation Army Daily. Retrieved from http://www.mod. gov.cn/wqzb/2016-01/28/content\_4637961.htm

Omohundro, S. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *AGI 171* (Vol. 171 of Frontiers in Artificial Intelligence and Applications).

Oren, M. (2009). UN report a victory for terror. *Boston.com*. Retrieved from http://archive.boston.com/bostonglobe/editorial\_opinion/oped/articles/2009/09/24/un\_report\_a\_victory\_for\_terror/

Orseau, L., & Armstrong, M. (2016). Safely interruptible agents. *Machine Intelligence Research Institute*.

Pistono, F., & Yampolskiy, R. (2016). Unethical research: How to create a malevolent artificial intelligence. *arXiv Preprint arXiv:1605.02817*.

Preliminary Statement of the Association of Manhattan District Scientists. (1945). The Gilder Lehrman Institute of American History. Retrieved from https://www.gilderlehrman.org/sites/default/files/inline-pdfs/03152.02\_FPS.pdf

Putin. (2017). Открытый урок «Россия, устремлённая в будущее». Retrieved October 28, 2017, from http://kremlin.ru/events/president/news/55493

Reed, S. R. (1990). Structure and behaviour: Extending Duverger’s Law to the Japanese case. *British Journal of Political Science*, *20*(3), 335–356.

Russell, S. (2017). 3 principles for creating safer AI. Retrieved from https://www.youtube.com/watch?v=EBK-a94IFHY

Sandberg, A., Armstrong, S., & Cirkovic, M. M. (2017). That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi’s paradox. *arXiv Preprint arXiv:1705.03394*.

Shackelford, S. (2009). From nuclear war to net war: analogizing cyber attacks in international law. *Berkeley Journal of International Law*, *27*(1).

Shulman, C. (2010). *Omohundro’s “Basic AI Drives” and Catastrophic Risks* (MIRI technical report). Retrieved from http://intelligence.org/files/BasicAIDrives.pdf

Shulman, C., & Armstrong, S. (2009). Arms control and intelligence explosions (pp. 2–4). Presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July.

Singh, S., Okun, A., & Jackson, A. (2017). Artificial intelligence: Learning to play Go from scratch. *Nature*, *550*(7676), 336–337.

Smith, P. D. (2007). *Doomsday Men: The Real Dr. Strangelove and the Dream of the Superweapon*. St. Martin’s Press;

Soares, N., & Fallenstein, B. (2014). *Aligning superintelligence with human interests: A technical research Agenda*. Machine Intelligence Research Institute (MIRI) technical report 8 (2014).

Tetlow, G. (2017, January 12). AI arms race risks spiralling out of control, report warns. Retrieved from https://www.ft.com/content/b56d57e8-d822-11e6-944b-e7eb37a6aa8e

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Tomasik, B. (2013). International Cooperation vs. AI Arms Race. *Foundational Research Institute*.

Torres, P. (2016, November 14). What a Trump Presidency Means for Human Survival: One Expert’s Take. Retrieved November 4, 2017, from https://medium.com/@philosophytorres/what-a-trump-presidency-means-for-human-survival-one-experts-take-ed26bf9f9a21

Trotsky, L. (1939). Three Conceptions of the Russian Revolution. Retrieved from http://www.wsws.org/en/articles/2008/10/rrev-o21.html

Turchin, A., & Denkenberger, D. (2017a). *Classification of Global Catastrophic Risks Connected with Artificial intelligence*. manuscript.

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

Turchin, A., & Denkenberger, D. (2017c). Levels of self-improvement. *Manuscript*.

Yampolskiy, R. (2015a). From Seed AI to Technological Singularity via Recursively Self-Improving Software. *arXiv Preprint arXiv:1502.06512*.

Yampolskiy, R. (2015b). Taxonomy of Pathways to Dangerous AI. Retrieved from https://arxiv.org/abs/1511.03246

Yudkowsky, E. (2002). The AI-Box Experiment. Retrieved from http://yudkowsky.net/singularity/aibox

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.

Yudkowsky, E., & Hanson, R. (2008). The Hanson-Yudkowsky AI-foom debate. In *MIRI Technical report*.

1. alexeiturchin@gmail.com [↑](#footnote-ref-1)
2. david.denkenberger@gmail.com [↑](#footnote-ref-2)