Classification of Global Solutions for the AI Safety Problem

Alexey Turchin

Science for Life Extension Foundation

alexeiturchin@gmail.com

David Denkenberger

Global Catastrophic Risk Institute (GCRI);

Tennessee State University;

Alliance to Feed the Earth in Disasters (ALLFED).

david.denkenberger@gmail.com

Abstract: There are two types of AI safety solutions: global and local. Most previously suggested solutions are local: they explain how to align or "box" a specific AI, but do not explain how to prevent the creation of dangerous AI. Global solutions are those that ensure any AI on Earth is not dangerous. The number of suggested global solutions is much smaller than the number of proposed local solutions. Global solutions can be divided into four levels: 1. No AI; AI technology is banned or its use is otherwise prevented; 2. One AI: the first superintelligent AI is used to prevent the creation of any others. 3. Net of AIs as AI police: a balance is created between many AIs, so they evolve as a net and can prevent any rogue AI from taking over the world. 4. Humans inside AI: humans are augmented or part of AI. We explore many ideas, both old and new, regarding global solutions for AI safety. They include changing the number of AI teams, different forms of the "AI Nanny" (non-self-improving global control AI system able to prevent creation of dangerous AIs), selling AI safety solutions, and sending messages to future AI.

Keywords: AI safety, existential risk, AI alignment, superintelligence, AI arms race.

Highlights:

- Global solutions of AI safety should be explored separate from local solutions
- AI alignment is a local, not a global, solution.
- Use of the first AI to take over the world is dangerous global solution
- Global solutions can be classified by the number of future superintelligent AIs: none, one and many.
- A potential AI arms race may be affected by regulating the number of the participants

Content:

1. Introduction

1.1. The AI safety problem

The problem of how to prevent of the global catastrophe connected with the expected development of AI of above human-level intelligence is often designated as

"AI safety" (Yampolsky & Fox, 2013). The topic has been explored by many researchers (Bostrom, 2014; Russell, 2017; Sotala & Yampolskiy, 2015; Yudkowsky, 2008).

An extensive review of possible AI safety solutions has been conducted by Sotala and Yampolskiy (Sotala & Yampolskiy, 2015). In their article, they explore classification of AI safety solutions as social, external, and internal measures.

In this article, we suggest a different classification of AI safety solutions, as *local* or *global*, and describe only global solutions. Local solutions are those that affect only *one* AI, and include AI ethics, AI alignment, AI boxing, etc. Global solutions are those that affect any potential AI in the world, like global technological relinquishment or use of the first superintelligent AI to prevent other AI from arising. Most solutions described by Sotala and Yampolskiy (2015) are considered local solutions in our classification scheme.

Recent significant contributions to the global solutions problem include Paul Christiano's model of slow takeoff (Christiano, 2018), which demonstrated that such takeoff could happen earlier than fact takeoff; Yampolskiy's research on AI arms races (Ramamoorthy & Yampolskiy, 2018), "Beyond MAD?: the race for artificial general intelligence"; Brundage's "The Malicious Use of Artificial Intelligence" (Brundage et al., 2018); and research on collective takeoff by Sotala (Sotala, 2017). The problem of "other AIs", central to the global AI safety conundrum, has been explored by (Dewey, 2016), who suggested four types of solution: international coordination, sovereign AI, an AI-empowered project, and some other decisive technological advantage.

Any local safety solution which cannot be applied globally will not affect course of human history, as many other AIs will appear with different local properties. However, some local solutions could reach the global level if an external transfer mechanism is added, such as an international agreement, or if the first AI based on this local solution becomes an only global power, *Singleton* (Bostrom, 2006).

In this article we will attempt classification of all known global AI safety solutions in hopes that it will help us to find previously unexplored solutions. We aim to make a strategic analysis of all possible solutions on the global level. We searched for suggested solutions in the literature and invented several of our own solutions.

We will classify solutions by the number of AIs involved. In addition, we will look at what group or groups could catalyze implement of such a solution, a local power (a state, or a foundation, or a group of people) or some global authority. As the world does not have any global governmental agency other than the currently weak UN, the solutions that require such agency are currently unworkable. After classification, we evaluate which solutions are most likely to be effective. The fact that we discuss each possible solution does not mean that we are advocating their application.

In Section 2 we overview the desired level of AI safety. In section 3 we look at solutions involving the prevention of the AI creation, while in section 4 we explore "One AI solutions", where the first AI prevents the appearance of other AIs. In section 5 we address many-AI solutions, in which many superhuman AIs appear and interact. In section 6, we suggest a class of solutions in which human beings exist inside AI.

In table 1 main ideas of the article are presented in the visual from.

Global Solutions of AI safety

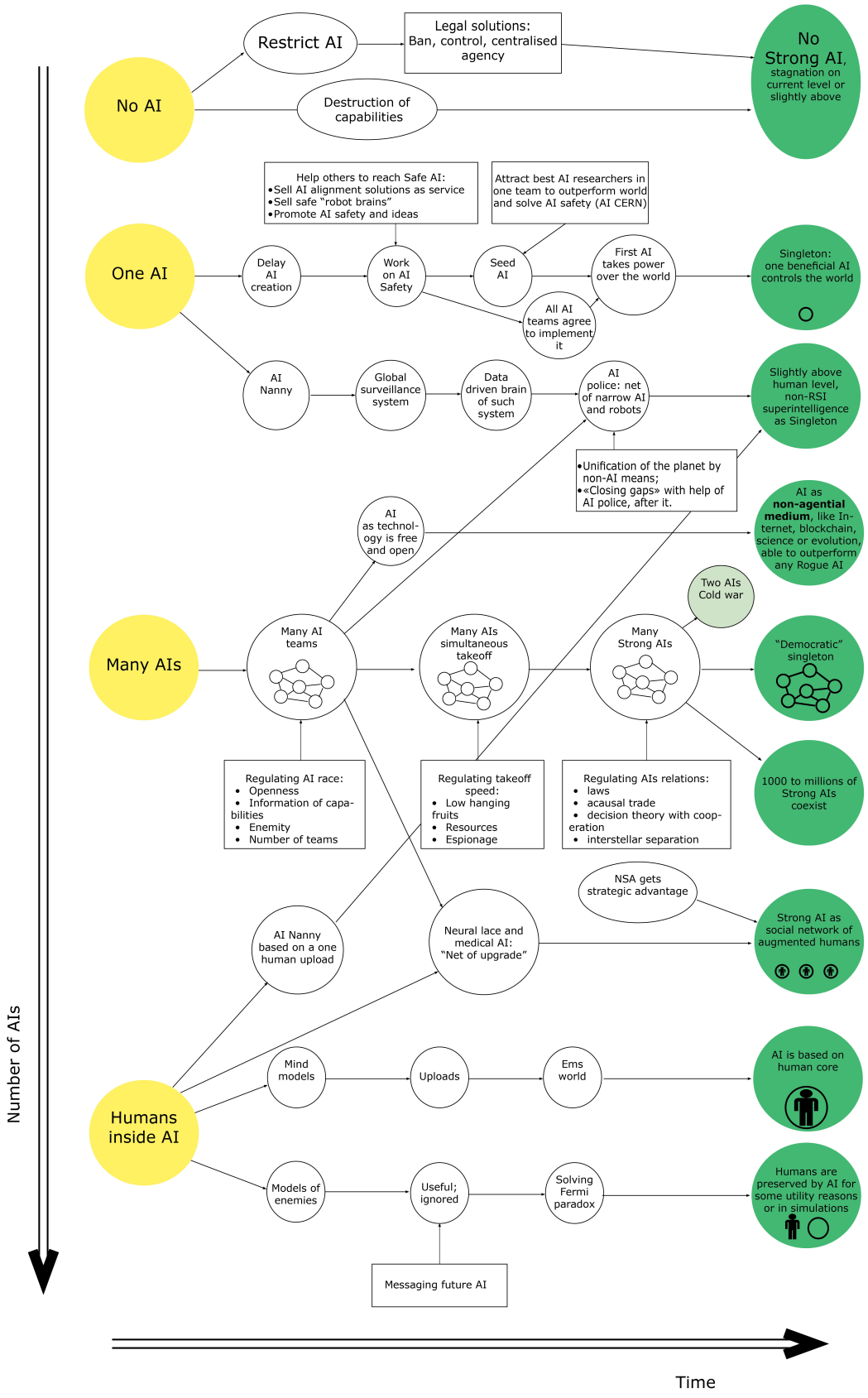

Table 1. Overview of the main ways to the global AI safety solution.

2. AI safety levels

To explore how to implement a global AI safety solution, we need some insight about what human safety may look like in the future. Global human safety in the far future (Beckstead, 2013) may be reached at different levels, from miserable survival to extreme flourishing. According Bostrom's classification, everything below full realization of the human potential is an existential risk (Bostrom, 2002), but low realization is not the same as extinction (Torres, 2016).

Several preliminary levels of AI safety may be suggested, similar to the classification of AI safety levels presented in a report from the Foundational Research Institute by Brian Tomasik (Tomasik, 2017), but centered on suffering. Our classification is based of levels of human well-being, the first and most basic of which is survival:

1) "*Lone wolf*:" At least one human being survives creation of strong AI, for example, as an upload.

2) "*AI survivors*:" A group of people survive and continue to exist after the AI creation, and may be able to rebuild human civilization to some extent. This may happen if the AI halts (Turchin & Denkenberger, 2018a) or leaves Earth.

3) "*No AI*:" Any outcome where global catastrophe connected with AI has not occurred. This is a world in which a comprehensive ban on AI is enforced.

4) "*Better now*:" Human civilization is preserved after AI creation in almost the same form in which it exists now, but benefits from AI in many ways, including curing of diseases, aging, crime prevention, material goods, interstellar travel, etc. Outcomes in this category involve a type of AI Nanny (Goertzel, 2012).

5) "*Infinite good*:" Benevolent AI will reach maximum possible positive utility for humans, but humans cannot now describe this utility as it is beyond our ability to imagine, as presented by Yudkowsky (Yudkowsky, 2004).

Different global solutions of the AI safety problem provide different levels of survival as the most plausible outcome. From our point of view, levels 3, 4 and 5 are acceptable outcomes, and outcomes 1 and 2 are unacceptable as they produce unimaginable human suffering and probably entail human extinction.

3. "No AI" solutions

In our world of quick AI development, AI relinquishment seems improbable and is essentially of purely theoretical interest. Many of these solutions have been explored by Sotala & Yampolskiy (Sotala & Yampolskiy, 2015).

Table 1. Overview of restrictive solutions

No AI	• International ban
	• Legal relinquishment
	• Technical relinquishment
	• Destruction of capability to produce AI
	  ○ War
	  ○ Luddism
	  ○ Staging small catastrophe
	• Slowdown of AI creation
	  ○ Economical
	  ○ Artificial AI winter
	  ○ Overregulation
	  ○ Brain drain
	  ○ Defamation of AGI idea

3.1. Legal solutions, including bans

Not many argue for a global AI ban as it is unfeasible under current conditions (Weng, Chen, & Sun, 2008) and would likely only help bad actors (Hughes, 2001).

One could imagine that global legal regulation could ban the creation of self-improving agents. But in our current, divided world its enforcement would be difficult. Only a powerful global government could make such a solution workable.

Some form of regulation may appear *ad hoc*, as an urgent measure implemented by the UN or the US plus a group of the most powerful countries. But they would need very credible harbingers as motivation. These could be several epidemics of AI-viruses of increasing strength, i.e. computer viruses with elements of machine learning (Turchin & Denkenberger, 2018a). However, there is currently no agreement what as to what factors would serve as a credible "alarm" for such harbingers, and such agreement may be impossible (Yudkowsky, 2017).

Some governmental and non-governmental groups are working to develop guidelines in this area. The EU is considering legislation about robotic ethics ("Robots: Legal Affairs Committee calls for EU-wide rules," 2017). Similar legislation may ban potentially dangerous self-improving systems, and if adopted in the most developed countries, it may act as a proxy for a global ban. It could be enforced in smaller, rogue countries by military coalitions, similar to the one formed in the 2003 Iraq war, but such a ban cannot be created and enforced without understanding the risks of AI. The recent Asilomar AI Guidelines (Future of Life Institute, 2017) could also serve as a foundation for internal control within the AI community to prevent creation of recursively self-improving (RSI) AI. The Asilomar guidelines could also form the basis for international law regulating AI.

Elon Musk recently advocated global regulation of AI research (Morris, 2017). Such regulations may take the form of a UN agency similar to the International Atomic Energy Agency (IAEA). The IAEA provides safety protocols for its members, demands openness and conducts inspections to confirm implementation; in exchange, it gives access to recent results on other members. The result will be something like Open AI, as described by (Brockman & Sutskever, 2015), but enforced by UN.

To implement such an AI agency, the UN would need a powerful enforcement agency. In the same way as when IAEA fails, international coalition uses sanctions (like against Iran and North Korea) or military intervention, like in Iraq. However, a UN-backed AI-control agency would require much tighter and swifter control mechanisms, and would be functionally equivalent to a world government designed specifically to contain AI. To be effective, such an agency should be empowered to use cyber weapons or even nukes. But in the current world climate, there will be little or no support for the creation of a world government authorized to nuke AI labs based only on theory. The only chance for its creation is if some spectacular AI accident happens, like hacking of 1000 airplanes and crashing them into 1000 nuclear plants by narrow AI with some machine learning capabilities. In such a case, a global ban of AI might be possible.

3.2. Restriction solutions

The idea of restriction is to find a scarce "commodity" needed for the creation of AI and try to limit access to it (Berglas, 2012). A global authority would be needed to implement such bans.

Such "commodities" could include:

- supercomputers
- programmers
- knowledge about AI creation
- semiconductor fabrication plants ("fabs")
- Internet
- electricity

The rarest commodity is chip fabs, which cost billions of dollars and are needed to create new processors. There are around 200 chip fabs in the world now ("List of semiconductor fabrication plants," 2017). If they were closed, no new computers could

appear in world, which might drastically slow AI progress. But the effect of fabs is rather indirect, as it is likely that enough computers already exist to create AI, especially given the large existing supply of graphic cards.

Large datacenters, supercomputers, scientific centers, and Internet hubs are also relatively rare, with the number worldwide in the thousands. Current home PCs (not connected to network) are probably unable to support AI, so if powerful computers and Internet connection are switched off, it could considerably slow down AI creation. These restrictions, like those discussed in section 3.1, would require preexisting global coordination. In addition, they would obviously have significant economic consequences.

3.3. Destruction solutions

One possible way to stop the creation of AI is annihilation by a nuclear attack of AI research centers, electronic equipment, and sources of electricity, which could be done locally or globally by a nuclear country acting alone. If such an attack was carried out against an adversary, it would "just" be a war; if done globally, it would mean that a superpower would bomb its own AI labs, which is unlikely. Nuclear attack of this type is extremely unlikely, unless it were perceived that an "AI uprising" had already started.

Destruction could be accomplished by a multitude of high-altitude electromagnetic pulses (HEMPs) caused by nuclear detonations. A concerted attack of this kind could destroy all unshielded electronics. Because electricity and fossil fuel extraction, and thus, industry, all depend on electronics, manufacturing and distribution would grind to a halt. This would not kill people directly, but could cause mass human starvation unless society were prepared (Cole, Denkenberger, Griswold, Abdelkhaliq, & Pearce, 2016), (Denkenberger, Cole, Griswold, Pearce, & Taylor, 2016). However, recovery of technological civilization and thus the ability to recreate AI is likely, so the problem would probably appear again. Alternatively, chaos could result

in a downward spiral leading to extinction. So, it is a risky "solution" that would likely only be temporary.

One could imagine other ways of destruction, ranging from economic recession to Luddism (Jones, 2013), to different global catastrophes, but all of them are impractical and morally unacceptable. In the future, some high-tech methods of AI halting may be implemented, like the Stuxnet computer virus that destroyed Iranian uranium centrifuges (Kushner, 2013). A virus may be used to destroy chip fabs, shut down the Internet, or cut electricity. There are other ideas in the field, but they are better not discussed publicly to avoid enabling the worst form of Luddite terrorists.

The unilateralist's curse – the lack of coordination between many actors with the same goal – (Bostrom, 2012) may exaggerate activities of those groups that least believe in the possibility of Safe AI. Prominent terrorist Ted "Unabomber" Kaczynski wrote a book in jail in which he claims the only way to prevent human extinction catastrophe is to organize a smaller catastrophe (Hanson, 2018).

3.4. Delay of AI creation

The global recession of 2008 did not have any measurable effect on the speed of AI development. Only a large-scale economic collapse that significantly disrupted global trade, could slow AI development to any significant extent.

Other events could slow down AI development, including:

- Fears of AI in the public.
- The next AI winter, lack of interest in its development (there have already been two after hype in the 60s and 80s).
- Extensive regulation of the field.
- Intentional creation of noise in the research field via fake news, defamation, white noise, and other instruments of informational warfare.
- Public ridicule of the field after some failure.

- Change of focus of public attention by substitution of terms. This happened with “nanotechnology”, which originally meant a powerful manufacturing technology, but now means making anything small. Such a shift may happen with the term “AI”, whose meaning has shifted recently from human-like systems to narrow machine learning algorithms. There are several fields that have had slow development for decades because of marginalization, such as cryonics, but it looks like the time of marginalization of AI has gone.

4. “One AI” solutions

These solutions are centered on the idea that the first AI will become dominant and prevent the development of other AIs. The nature of these solutions is that they are implemented locally but affect the whole globe.

Table 2. Overview of “one AI” solutions

One AI	• First AI is used to take over the world o First AI is used as a military instrument o First AI gains global power via peaceful means - Commercial success - Superhuman negotiating abilities o Strategic advantage achieved by narrow AIs produces global unification before the raise of superintelligent AI by leveraging preexisting advantage of a nuclear power and increasing first-strike capability o First AI is created by a superpower and provides it a decisive strategic advantage o First AI is reactive, and while it doesn’t prevent the creation of other AI, it limits their potential danger

	○ First AI is a genius at negotiation and solves all conflicts between other agents • First AI appears as a result of collective efforts ○ AI police: global surveillance system to prevent creation of dangerous AI ○ AI CERN: international collaboration creates AI Nanny ○ Main players collaborate with each other ○ AIs are effective in cooperation and merge with each other • Non-agential AI medium (AI as widely distributed technology, without agency) ○ Distributed AI based on blockchain (*SingulatityNET*) ○ AI as technology everywhere (openness) ○ Augmented humans as AI neurons (*Neuralink*) ○ Superintelligence as a distributed optimization process of concurrence between AI agents (*market*) **Indirect measures to increase probability that first AI will be human-aligned:** • Helping others to create safe first AI ○ AI safety theory is distributed among main players and used by every AI creator ○ AI safety instruments are sold as a service ○ Promotion of AI safety • Slowing creation of other AIs ○ Concentrate best minds on other projects and remove them from AI research ○ Take low-hanging research fruit

	• Factors affecting the arms race for AI include funding, openness, number of teams, prizes, and public attitudes • Sending messages to any future AI

3.1 Create the first AI and use it to take world power

AI could potentially take world power on its own. Max Tegmark describes a scenario in which the first AI initially gains world dominance through earning money and later consolidates power by rigging elections or staging coups in different countries (Tegmark, 2017).

The main problem of the idea that first AI can be used as an instrument to take over the world is that it creates motivation for militarisation of AI, which has potentially dangerous consequences (Turchin & Denkenberger, 2018b).

Superintelligent AI may be able to find win-win solutions in negotiations. Such an ability could help it overcome resistance to global unification, as it will be able to provide its unique negotiation ability as a service, to which everybody will be interested to apply, and in that case, there will be no need for military world takeover.

3.1.1. Concentrate the best AI researchers to create a powerful and safe AI first

This idea is to create something like the Manhattan Project, attracting the best minds to work together on the creation of the first self-improving AI. This would provide such a large concentration of human intelligence that they could simultaneously create AI and solve the problem of AI safety. The Manhattan Project was formed of the best scientists in the world, and they were concerned about potential global risks of the first nuclear explosion. Scientists involved in the project created the LA-602 (Teller, 1946) report about the possibility of nuclear chain reactions in the atmosphere.

Later efforts to create nuclear weapons in other countries were not so safety-oriented. The Soviets exploded a bomb over their own troops (Ria Novosti, 2009). The

Indians dropped explosives intended to be part of their first nuclear bomb during critical assembly—fortunately, it did not detonate (Nuclearweaponarchive, 2001).

If a similar trend holds for AI research, the first concerted effort may be more safety-oriented and involve better planning and brighter minds than later efforts. In addition, if research is accelerated in one research institution, it could outperform the world in general. This could help prevent a troubling situation in which safety solutions are well-understood in one organisation but AI is created by another company.

If the first effort is ahead of the competitors by years, it will have a safety time gap, that is additional time for working on AI safety. In current situation, it looks like Google has the necessary advantage and could be ahead of competitors for 1–2 years.

In early stages of its development (in the 2000s), the Machine Intelligence Research Institute (MIRI) had a plan to be the creator of the first Friendly AI. However, its goal now is to facilitate research on AI safety solutions (MIRI, 2016), ("About MIRI," 2017) to be implemented elsewhere.

3.1.2. Using the decisive advantage of non-self-improving AI to create an AI Nanny

Sotala (Sotala, 2016) wrote that even non-self-improving AI may gain a decisive strategic advantage if it is effective at designing new weapons, or in strategic games of political planning. This opens the possibility to use the first human-level AI to gain power over the world, without taking the dangerous and unpredictable route of recursive self-improvement.

Such AI may be built around a human upload, which gains most of its power not from self-improvement, but from running on high-speed hardware. Such a high-speed human upload gaining global power via social manipulation and designing new weapons may become an "AI king".

One way to such gain a decisive strategic advantage is via a first AI created by a superpower (either China or US) which is already close to world domination. Such an AI, created as a government-sponsored large project, may be attained as part of a secret "Manhattan Project"-type effort or by seizing the archives and work of a large private company. The AI could leverage other power-projecting instruments already controlled by this superpower to provide it with the capability for world domination (like access to secret information, control of nuclear weapons, large financial resources). See recent remark of Putin that "the nation that leads in AI 'will be the ruler of the world' (Putin, 2017).

For example, even narrow AI designed to calculate nuclear war scenarios could provide a decisive strategic advantage for an existing nuclear superpower. It could then strike in the way that there is a high probability of no retaliation.

Dewey (2016) suggested first AI may be reactive and proactive: Proactive AI prevents creation of other AIs, starting preemptive wars against them, and reactive AI only limits or ensures the safety of other fast AI take-offs. Dewey also suggests that two types of strategic advantage, proactive or reactive, may be reached by non-self-improving AI. In his opinion, another option is strategic advantage reached by non-AI technological means.

3.1.3. Risks of creating hard-takeoff AI as a global solution

In AI safety research it is typically assumed that the first superintelligent AI will take action to prevent the creation of other AIs. In that case, solving local AI safety would provide global safety.

However, if the first AI is created in, say, the US, it must prevent creation of another AI in China. From the point of view of international law, an action AI takes in this direction would be an act of war (Turchin & Denkenberger, 2018b).

Deliberately creating an AI that will start a war immediately after its creation is very provocative for other actors. In the face of such a threat they might use a

preemptive nuclear strike to prevent the creation of AI. Kahn (Kahn 1959) wrote the same of the potential creation of a Doomsday nuclear bomb that could kill all humanity—that just the act of its creation could be even more provocative than a nuclear attack.

Not just the actual creation, but just the intention to create such AI, may attract attention from foreign and domestic secret services. Publicly suggesting that the first creators of AI should program it to take over the world may have legal consequences (as such AI is illegal cyberweapon) and may prevent open dissemination of any AI safety theory based on such suggestion.

It appears that creation of a military infrastructure is a convergent instrumental goal for any first AI (Turchin & Denkenberger, 2018b). This infrastructure would help the AI prevent creation of other AIs as well as prevent humans and government agencies from trying to switch off the AI. If other AIs are in advanced stages of development, they will resist the attempt to shut them down. In this case, a war between AIs will start, in which humanity could perish or be taken hostage. Thus, this solution is intrinsically risky and better solutions should be sought.

Another idea is that the creation of AI safety theory will happen separately from the creation of AI, but the first AI creator will use available safety theory. We will discuss this possibility in section 3.3.

3.2. One global AI created by collective efforts

3.2.1. AI Nanny requires a world government for its creation

The idea of an AI Nanny has been suggested by Ben Goertzel, who has described "…the creation of a powerful yet limited Artificial General Intelligence (AGI) system…with the explicit goal of keeping things on the planet under control while we figure out the hard problem of how to create a probably positive Singularity. That is: to create an 'AI Nanny.'" (Goertzel, 2012)

He proposed the following properties for an AI Nanny:

- General intelligence somewhat above the human level,
- Interconnection with powerful worldwide surveillance systems,
- Control of a massive contingent of robots, and
- A cognitive architecture featuring an explicit set of goals.

Muehlhauser and Salamon (Muehlhauser & Salamon, 2012) criticized this idea because solving AI safety for the AI Nanny would require solving almost all AI safety problems for self-improving AI.

The AI Nanny also does not solve the main problem of how the first AI will gain its global power—by world takeover or by peaceful integration of a net of AIs. The first way has its own risks and the second could have dangerous holes. One possible solution here is peaceful integration of most of the world, plus than forceful integration of remaining "rogue states". This would resemble the current fight of a large international coalition that already owns nuclear weapons with "rogue countries" that try to make their own nuclear weapons.

A united world government may be required for the creation of an AI Nanny, but under current conditions, such a world government is unlikely to peacefully appear. Such a world government may appear if one country gains an overwhelming military advantage from a means other than AI. If the advantage arose from AI, the problem of AI safety would already be solved, but it could come from powerful nanotechnology weapons or some type of narrow-AI robotics. Alternatively, if the risks of AI are highly visible, or perhaps already felt, most countries may give up their sovereignty to the UN to create an AI Nanny. Such a scenario could happen if a narrow-AI-based computer virus created widespread devastation of infrastructure, or if the first self-improving AI appears but spectacularly fails at some stage of its development.

The AI Nanny may have rather high intelligence, but in a form which is not easy to self-improve, e.g., a large database of pre-recorded solutions and neural

algorithms, as well as all existing data about the world and from surveillance systems. Such a "data-driven" AI may be a relatively safe local solution.

Some semi-universal AI may be created in the current age of neural nets (Paul Christiano, 2016) as a very large and prohibitively expensive international project, like the Human Genome Project, Large Hadron Collider and International Thermonuclear Experimental Reactor. Gary Marcus recently suggested that we need something analogous to the European Organization for Nuclear Research, CERN, for AI (Itut, 2017), in a sense similar to Baruch's 1946 plan to centralize nuclear research (Dewey, 2016).

An AI Nanny could be designed on many opaque neural net modules that would prevent its self-improvement, and its enormous size would prevent it from leaking into the Internet. Its intelligence also may not be universal or not exceed total human intelligence. Thus, an AI Nanny would likely be rather safe and under international control. But it looks like the opportunity for such a project is lost, as many large companies are now participating in their own projects and there is a lot of available hardware as well as openly published materials. The potential is not yet completely lost; large international collaborations like "Partnership for AI" (Partnership for AI, 2017) may be similar to an AI Nanny.

3.2.2. Levels of implementation of the AI Nanny concept

We suggest 4 levels of possible intelligence of an AI Nanny:

1. Use of a distributed surveillance system, which does not have much intelligence but is able to enforce a universal ban on creation of self-improving systems. This is a low-level solution.

2. Creation of neural-net-based and data-driven AI as part of a large international project. In this case, the AI's intelligence comes not from fluid intelligence but from extensive knowledge and models. It may serve as the brain of the surveillance

system mentioned above. One possible solution is to use first human upload as an "AI king", or world governor, with the main mission of preventing the creation of other AIs (Turchin, 2017a). An AI king will run at higher speeds than ordinary humans, using all available hardware, which will give him greater intelligence while maintaining alignment with human values. The idea is controversial from technical, political, and moral points of view.

3. Creation of AI police, a net of narrow AIs able to control the appearance of self-improving AIs and other dangerous entities.

4. Creation of high-intelligence AI Nanny as described by (Goertzel, 2012). This AI will be as much above humans, as humans are above apes; it would be some form of superintelligence (SI). In that case, we would have exactly the same problems as with control of any other strong AI (Muehlhauser & Salamon, 2012). But if we create a weaker system, we could probably find Goldilocks' path between its ability to control research and our ability to control the system.

3.2.3. Global transition into AI: non-agential AI-medium everywhere, accelerating smoothly without tipping points

The AI described above was agential. But some of the strongest known optimization processes are non-agential: e.g. evolution, market forces, and science. These processes appear from the interaction of millions of agents with their own goals, and the optimization power of these processes does not depend much on direct summing of the minds of agents. Instead, it is a result of their interactions, so it is not a net of AIs, which will be discussed below, as net implies higher level of goal's coordination.

The AI medium self-improves more quickly than any individual part of it, because self-improvement is a property of the whole system, but not of any one part of it, as it results from the way information is exchanged between different parts.

We will call such processes "intelligent media", as opposed to intelligent agents, as they do not have independent goals but perform any tasks they find. This medium is a form of environment; as such, it does not conquer territories but attracts other agents to participate in it; a similar idea has been suggested by Mahoney (Mahoney, 2008). This feature could still be devastating, as we know that in an analogous case, market forces can destroy traditional cultures more effectively than weapons (Alexander, 2016). A non-agential AI medium does not have to take over the world because it would simultaneously appear everywhere.

It would not be surprising if superintelligence also arises from a medium. This idea in naïve form has been presented as "the Internet will gain consciousness". The Internet surely will be a backbone for the AI medium, but something more is needed. One can imagine other elements of an AI medium in the form of blockchain, social networks, prediction markets (Hanson & Sun, 2012) and scientific references net (Camarinha-Matos & Afsarmanesh, 2005). One of the routes to an AI medium is to connect all human brains through some form of net, producing a global brain (Luksha, 2014).

There are concerns that such collective evolution is unstable and will eventually produce one agent that will be able to improve itself more quickly than the overall AI medium and thus destroy it. See the latest review of Kaj Sotala about difference between individual and collective takeover (Sotala, 2017), criticizing Vinding's recent book (Vinding, 2016).

"AI Foom" is an intelligence explosion (Yudkowsky, 2006), and most arguments for "collective foom", that is, self-improving AI medium, try to prove that it will be a natural course of events, but is it a preferable outcome and could its probability be increased? Scott Alexander showed that it is possibly a negative outcome in the form of ascending economy (Alexander, 2016), which destroys all human values to increase

market output. The market has been criticized by Marx (Marx, 1867) for the same flaws..

As the AI medium evolves without values, it cannot be friendly or unfriendly to humans, so it will only prioritize human survival, if humans will be able to trade with it, or be ignored by it.

John Smart (Smart, 2012) predicted that the evolution of such a system will consist of constant acceleration and miniaturization, which could be described by a hyperbolic law.

3.3. Help others to create safe AI

3.3.1. Promoting ideas of AI safety in general and the best AI safety solution to all players

As we mentioned above, it is *a priori* improbable that the same team that creates optimal AI safety theory will also create the first AI unless it is part of a CERN-style giant collaboration. Thus, the AI safety team should try to make other AI teams adopt AI safety theory they created.

There are a number of tangential measures that may help in the development of AI safety, but do not guarantee results, including:

- Funding of AI safety research.
- Promotion of the idea of AI safety.
- Protest military AI.
- Friendly AI training for AI researchers.
- Provide publicly available safety recommendations.
- Increase "sanity waterline" and rationality in the general population and among AI researchers and policymakers.
- Lower global levels of confrontation and enmity.
- Form political parties for prevention of existential risks and control of

AI risks. However, even if such parties were to win in larger countries and able to change policy, there would still be small or undemocratic countries that could use technology freeze in larger countries as an advantage.

Another idea is to seek ways to attract the best minds to solve the AI safety problem. Yudkowsky said that one of reasons he wrote the book HPMOR (Yudkowsky, 2010) was to attract the best mathematical minds to the AI safety problem. Attracting top minds would achieve simultaneously several useful goals:

- Depleting the pool of minds for direct—not necessarily safe—AI research, thereby slowing it down
- Increasing the preponderance of AI safety theory
- Establishing relation with the best AI teams, as some of people who worked on AI safety may eventually join such teams or may have friends there, and
- Promoting the idea that unlimited self-improvement is dangerous and unstable for all players, including possible AIs.

3.3.2. Selling AI safety theory as an effective tool to align arbitrary AI

If AI safety theory can align the goals of an arbitrary AI, it will be very attractive for any reasonable AI creator, as the creator insures their own safety and ability to instil goals to the AI. The AI creator could save many resources by implementing a proven alignment method. However, while it lessens the probability that the AI will run amok, the creator still could align the AI with a dangerous, egoistic goal.

If an AI safety tool-kit could be sold as a good, this would increase the likelihood that first movers will use it, as it would be profitable for them. It could also be sold as a service, which could include custom adaptation and training. Selling "AI safety" may produce a wider reach than just publishing a pdf with explanations, and the customer relationship could increase its implementability.

3.4. Local action to affect other AIs globally

3.4.1. Slowing the appearance of other AIs

The idea is that people could take some actions locally that will affect any other AI globally, which may appear in the future at an unknown location.

Such actions may include espionage or *taking low-hanging fruit in research*, which will increase overall level of the technology, but lower chances that one of the participants of the race will leapfrog others by taking such low hanging fruit; draining the pool of easily available resources, which includes both minds and hardware, may also be regarded as taking low hanging fruits. While it is impossible to drain all hardware, the leader in AI research could invest in owning leading positions in hardware capabilities as well as training datasets for neural nets.

3.4.2. Ways to affect a race to create first AI

An AI creation race is generally regarded as bad because it encourages creation of the least-safe AIs first. A war between AIs may also become possible if several AIs are created simultaneously (Bostrom, Armstrong, & Shulman, 2013; Shulman, 2011).

There are many ideas how to affect the AI race to make it safer, that is to lower the probability of creating dangerous AI. As a race with many participants is a very complex game, there are not obvious ways to predict how it will react to seemingly good interventions, like openness. Bostrom has shown (Bostrom, 2016b) that if nobody knows the capabilities of others and their own capabilities, it will slow down the race, so openness about capabilities may be dangerous.

Actions that may affect an AI race and make it safer may include:

1) Changing the number of participants (more on it in section 4.2.2).

2) Increasing or decreasing information exchange and level of openness.

3) Reducing the level of enmity between organizations and countries and preventing conventional arms race and military buildup.

4) Increasing the level of cooperation, coordination, and acceptance of the idea of AI safety among AI researchers.

5) Changing the total amount of funding available.

6) Promoting intrinsic motivation for safety. Seth Baum discussed the weakness of monetary incentives for beneficial AI designs, and cautions: "One recurrent finding is that monetary incentives can reduce intrinsic motivation" (Baum, 2016); when the money is gone, people lose motivation. Baum also noted that the mere fact that a law existed promoted obedience in some situations and that social encouragement can increase intrinsic motivation.

7) Changing social attitudes toward the problem and increasing awareness of the idea of AI safety.

8) Trying to affect the speed of the AI race, either slowing it down or accelerating it in just one place by concentrating research. It is interesting to note that acceleration could be done locally, but slowing it would require global cooperation, and so is less probable.

9) Affecting the idea of the AI race as it is understood by the participants (Baum, 2016): if everybody thinks that the winner takes everything, the race is more dangerous. A similar framing solution has been suggested in the field of bioweapons, that is, to stop claiming bioweapon creation is easy, as it may be attractive to potential bioterrorists. In fact, bioweapons are not as easy to develop and deploy as is shown in movies, and would probably kill the terrorists first (Ouagrham-Gormley, 2013).

10) Affecting the public image of AI researchers who are currently presented as not wanting beneficial AI design (Baum, 2016).

11) Refraining from suggestions of draconian surveillance as they "inadvertently frame efforts to promote beneficial AI as being the problem, not the solution" (Baum, 2016).

12) Stigmatization of building recursive self-improving AI by framing them as morally unacceptable, as was done with landmines. Stigma impelled even countries that did not sign the treaty that prohibits landmines to reduce production (Baum, 2016).

13) Deliberate association with crackpottery: an example is UFO research: anyone who mentions the word "UFO" will no longer be accepted in the scientific community as a credible scientist. It probably worked for AI during AI winters, when scientists tried not to mention the words "artificial intelligence". Society could associate "self-improving AI" with craziness, which would be not difficult if we pick some of the most outstanding ideas from associated Internet forums, like *Roco Basilisk* (Auerbach, 2014). Such an association may reduce funding for such research. However, AI could start to self-improve even if it was not designed to do so; thus, such association would probably be damaging to AI safety efforts. Recent successes in meta-learning of neural nets by Google's *DeepMind* show that the idea of self-improving AI is becoming mainstream (Fernando, 2017).

14) Affecting of the speed of takeoff after one AI starts to win. If the speed of self-improvement of one AI diminishes, other AIs may catch up with it.

We address some of these ideas in the next section.

3.4.3. Sending messages to future AI or participating in acausal deals with it

Rolf Nelson (Nelson, 2007) suggested that we could install indexical uncertainty into the future AI; in that case, if we make a commitment now that if humanity creates a friendly AI, this friendly AI will also create simulations of most probable types of rogue AI, which will be turned off if a given AI does not simulate benevolence to

humans. In that case, any rogue AI will be uncertain if it is in a simulation or not, and as killing humans has small marginal utility in most cases, it would prefer emulate benevolence. Turchin has explored this and several other similar moves in detail (Turchin, 2017b). But such an approach would probably work only for AI Singleton, and it is our last level of defense.

4. "Many AI" solutions

4.1. Overview of the "net solutions" of AI safety

4.1.1. How a net of AIs may provide global safety

In a nutshell, the idea of a "net solution" to AI safety is that there will be many AIs, and this fact will provide some form of protection. The most prominent backer of this approach is Elon Musk, who wants to unite AI working teams in a net based on openness and upgrade humans, so they will not obsolete in the age of AI (Kharpal, 2017). However, there are risks to this approach (Bostrom, 2016b).

There are two main features, which may provide safety from a net of AIs:

1) **Intelligence overhang of many AIs (the net of AI has much higher intelligence than any rogue AI, so it is able to create effective protection) and the net as self-adapting surveillance system.** An AI-net forms something like AI police, which prevents any single AI from unlimited growth. This is analogous to the way 30 trillion human cells provide a multilevel defense against unlimited growth of a single cell—cancer—in the form of an immune system. The approach is somewhat similar to the AI Nanny approach (Goertzel, 2012), but an AI Nanny is a single AI entity. An AI-net consists of many AIs, which use ubiquitous transparency (Brin, 1998) to control and balance (Hanson, 2016) each other.

2) **Value diversity** of many AI sovereigns (Bostrom, 2014), (Bostrom, 2016a) guarantee that different positive values will not be lost. Different members of the net have different terminal values, thus ensuring diversity of values.

We will call this many-AI solution a "Multipolar Singleton" (Bostrom, 2006), as global coordination will result from constant negotiation and trade between entities with different values. A Multipolar Singleton will have the following necessary conditions:

- Many superhuman AIs exist.
- The AIs all find mutual cooperation beneficial, and have some mechanism for peaceful conflict resolution.
- The AIs have diversity of final goals, so some goals are more beneficial to humans than others. This protects against any critical mistake in defining a final goal, as many goals exist. However, it is not optimal, as some of AIs may have goals that are detrimental for humans. It will be similar to our current world, with different countries, but the main difference will be that they will likely be much better able to peacefully coexist than currently, because of AI support.
- The fact that Earth is surrounded by infinite space. Different AIs could start travel to stars in different directions, and as each direction includes infinitely many stars, even "infinite goals" could be not mutually exclusive and may not provoke wars.
- Finding it mutually beneficial to create something like an immune system or AI police to prevent unlimited self-improving AIs or other dangerous AIs from developing via ubiquitous intelligent control.

The main question is how to reach an AI-net solution and whether it will be stable, collapse into war between AIs, or consist of one AI dictatorship.

4.1.2. Different sizes of a net of AIs

The most important variable here is the number of future superintelligent AIs, which depends on the speed of AI self-improvement and the number of teams of AIs.

The slower the AI takeoff is, the larger the number of AIs will be, though this also depends on the number of AI teams, among other factors. There are several vague groups of the possible numbers of coexisting superintelligences, which will have different dynamics, including:

1) Two AI sovereigns' semi-stable solution, like the Cold War (Lem, 1959).

2) From several to dozens of sovereign AIs, similar to existing nation-states; they may be evolved from nation-states, or from large companies.

3) From thousands to billions AIs, with relations similar to relations between humans now, probably resulting from some brain uploading technology (Hanson, 2016), human augmentation (Urban, 2017), or genetic modification (Bostrom, 2003b).

4) Uncountable or almost-infinite number of AIs, similar to AI-medium, discussed above. This could be similar to IoT, but with AIs as nodes.

In the table 3 we present overview of possible solutions, which will be explored in detail below.

Table 3. Net of AIs

Many AIs	• Net of AIs forms a multilevel immune system to protect against rogue AIs and has a diversity of values, thus including human-positive values o Increase number and diversity of AIs - openness - slowdown of AI growth - human augmentation

	- self-improving organizations - increase number of AI teams - create many copies of the first AIs o Uploads come first • Several AI-sovereigns co-exist, and they have better defensive than offensive capabilities o Two AIs semi-stable "Cold war" solution - tight arms race - military AI evolution - MAD defense o AI-sovereigns appear from nation states - very slow takeoff and integration with governments o AIs expand in space in different directions - creation of AIs on remote planets

4.2. From the arms race between AI creating teams to the net of AIs

4.2.1. Openness in AI development

Elon Musk and others presented the idea of OpenAI in 2015: "We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as is possible safely" (Brockman & Sutskever, 2015).

In the following, we discuss the idea of openness of the field of AI and the net of AIs as we understand it; it does not necessarily represent the position of the "OpenAI" initiative.

We look at the following approach: many AI projects freely exchange ideas, datasets and progress results, thus accelerating AI creation and ensuring its safety. We will call it an “open net of AI teams”.

Safety emerges from the following characteristics of such collaboration:

- None of the AI teams gains strategic advantage over other teams, as all the data from every team’s results are available to all of the teams. An attempt to hide results will be seen publicly. Openness ensures that many AI teams will come close to self-improving AI simultaneously, and there will be many such AIs, which will balance each other.
- The teams outside the “open net” are much less likely to gain strategic advantage, as they are not getting all benefits of the membership in the net, namely access to the results and capabilities of others. But this depends on how much information becomes part of the public domain. Open net will have an “intelligence advantage” over any smaller player, which makes it more likely that self-improvement will start inside the open net, or that the net will have time to react before the rogue agent “outsmarts” the net.
- The “open net” will create many different AIs, which will balance each other, and probably will be motivated to engage in mutually useful collaboration (However, some could take advantage of openness of others but not share their own data and ideas). If one AI leaves the net for uncontrolled self-improvement, the collective intelligence of the net will still be higher than that AI for some time, probably enough to stop the rogue AI.
- The value system of the net will provide necessary diversity, so many possible goals will be presented to at least some extent. This lessens the

chance that any good goal will lost but raises the chance that some AI projects will have bad, dangerous, or otherwise unacceptable goals.

- Because of their ability to collaborate, the "open net" may be able to come to unanimous decisions about important topics, thus effectively forming a Singleton.
- The net will be able to take all low-hanging fruits of self-improvement, like the ability to buy hardware or take over the Internet, thus slowing down self-improvement of any rogue agent.
- The net will help to observe what the other players, not involved in open net, are doing—for example, the fact that German scientists stop publishing articles about uranium in 1939 showed that they were trying to keep their work secret and hinted that they were working on a bomb.
- The net will contribute to the creation AI vigilantes or AI police, as suggested by David Brin in his transparent society proposal (Brin, 1998). Therefore, the open net may somehow evolve in the direction of an AI Nanny, perhaps consisting of many distributed nodes.

Bostrom has criticized the idea of openness in AI, because he feels it could accelerate dangerous research (Bostrom, 2016b). It would also not be easy to balance a dangerous AI, as it could undertake local actions that could immediately kill everybody, like constructing a very large nuclear cobalt bomb (Smith, 2007) or a dangerous biological virus. But if there are many AIs, they probably could have the needed level of mutual control to prevent local dangerous actions or contain the results of such actions.

The main question is if openness in AI will be able to prevent a rogue actor from using these data to start self-improving first and gaining a decisive advantage over

others. These worries are described in an excellent post by Scott Alexander (Alexander, 2015).

Above we assumed that net of teams will create the net of AI, however, the net of teams may cooperate in creation just one AI.

4.2.2. Increase of the number of AI labs, so many AIs will appear simultaneously

Bostrom explored the situation of many competing teams depending of their number, their enmity, and their knowledge about their own and each others' capabilities. He found that the fewer the number of the teams, the smaller the overall risk, and also it is better if they do not know about each other or about their own capabilities (Bostrom et al., 2013).

In fact, there are already many AI teams and such a large number may result in many simultaneous AI takeoffs. History shows that some important discoveries were made independently with a very small temporal separation. For example, the first two telephone patent applications were filed within 3 hours of each other on February, 14, 1876 (Baker, 2000) and the Soviet–US race to bring material back to Earth from the Moon was decided by 3 days in 1969 (The Telegraph, 2009).

Increasing the number of independent AI projects will increase probability that several of them will have hard takeoff simultaneously, but it will also the increase chances that some of the programs will have a very low level of safety, as Bostrom noted (Bostrom et al., 2013). Increasing the number of teams is a local action that has global consequences.

The publicity around AI in recent years has likely contributed to the growth of AI companies. Venture Scanner tracked 957 AI-creating companies (Venture scanner, 2016). While most of them are not building real AGI, many of them would be happy to have as AI as universal as possible. It is also clear that many companies and

individuals are not presented in this list, including university projects and individual researchers. It could also be that some companies in the list are fake or should not be counted for other reasons. So, it is reasonable to estimate the total number AI teams now working as within an order of magnitude of 100, but most of the research is coming from around ten major companies including Google, Facebook, and Open AI.

This means that there may be no need to increase the number of teams to prevent a single dominant AI—their number is already on the order of magnitude where several hard takeoffs could happen simultaneously.

4.2.3. Change of the self-improving curve form, so that the distance between self-improving AIs will diminish

Yampolskiy showed that there are several reasons why the actual self-improving of one AI system may be described by a logarithmic rather than exponential curve (Yampolskiy, 2015b). However, artificial interventions like taking low-hanging fruits and espionage could change this rate. If the curve is shallower, more AIs will reach the level of superintelligence simultaneously, providing a better chance for some balance of power.

4.3. Instruments to make the net of AIs safer

4.3.1. Selling cheap and safe "robotic brains" based on non-self-improving human-like AI

This idea is to make a safe AI design, which can solve almost all tasks that other people or organizations may need. Such a design would then be provided widely and very cheaply either as hardware or from the cloud. This would undermine the economic need for creation of other AIs and create the opportunity for a global AI Nanny. This non-self-improving, safer AI is analogous to the idea of non-self-replicating safer

molecular manufacturing, like a nanofab, which is regarded a safer form of nanotech than nanorobots (Drexler & Phoenix, 2004).

One possible design of such a "robotic brain" may be a human upload (Hanson, 2016) or some simplified model of a human brain, which finds a balance between upload and neuromorphic AI (Turchin, 2017a).

4.3.2. Starting many seed AIs simultaneously

Yudkowsky suggested (Yudkowsky, 2001) that the first self-improving AI will appear from a seed AI, which is a relatively small program. This program may even start out with a below-human level in terms of its intelligence and capabilities, but be able to recursively improve its own code and thus quickly gain immense power.

One way to balance such a system is to start many seed AIs with slightly different initial conditions, so they will form an ecosystem with different values. Such an approach will likely have unpredictable consequences, and may be used only as a backup measure, if control over the first seed AI is lost. This is applicable to any AI—if the control over it has been lost, another copy of the same AI could be started from the backup.

5. Solutions in which "Humans are incorporated inside AI"

5.1. Different ways to incorporate humans inside AI

Some form of superintelligence may be created with humans as participants in it. However, as Bostrom shows (Bostrom, 2014), there is always the problem of the "second transition", that is, the appearance of more powerful AI inside such a system,

which no longer needs humans. So, any such system must create AI police to prevent "second transition".

Another problem is that most such solutions are lagging, as human uploading is still far away. "Self-improving organizations", organizations investing in multilevel optimization of their structure, from building custom hardware to corporate culture—the best examples here are *Alphabet* and Musk's business empire—have recently had the biggest success in AI development.

Some ways of incorporating humans inside AI include:

1. AI could be built around a human core or as a human emulation. It could result from effective personal self-improvement via neural implants (Bostrom, 2014), adding tool AIs and exocortex. There is no problem of the "AI alignment", as there are no two agents that should be aligned, but only one agent whose value system is evolving (Turchin, 2017a).

2. AI could appear from a net of self-improving posthumans, connected via neural interfaces (Sotala & Valpola, 2012). This combines ideas of social networks, blockchain, and Neuralink (Urban, 2017). Such a net may appear from the evolution of medical AI (Batin, Turchin, Markov, Zhila, & Denkenberger, 2018).

3. AI could result from genetic modification of humans for intelligence improvement (Bostrom, 2003b).

4. Superintelligence could appear as a swarm intelligence of many human uploads and not evolve in a more effective and less human form for some unknown reason (Hanson, 2016; Scott, 2016).

5. Only one human upload is created, and it works as an AI Nanny, preventing the emergence of any other superintelligences (Turchin, 2017a).

6. Superintelligence is created by a "self-improving organization" as a property of the whole organization, which includes employees, owners, computers, hardware-building capabilities, social mechanisms and owners. It could be a net of self-

improving organizations, similar to Open AI (Brockman & Sutskever, 2015) or "Partnership for AI".

7. Nation-states evolve into AI-states, and keep most of their legislation, structure, values, people, and territories. This is most probable in the case of the soft takeoff scenarios, which would take years. Earth could evolve into a bipolar world, similar to the Cold War. We should expect a merger between self-improving organizations and AI-states, perhaps by acquisition of such companies by state players.

Is not easy to envision them now, but there could also be a scenario that combines some of the directions listed in this section.

5.2. Even unfriendly AI will preserve some humans or information about humans

Below is an assortment of less-probable ideas that generally provide a lower level of safety (levels 1 and 2). In these scenarios, human beings will somehow be incorporated, used, or remembered by unfriendly AI.

- Unfriendly AI may have a subgoal to behave as benevolent AI toward humans, based on some Pascal mugging-style considerations and ontological uncertainty if it will think that there is small chance that it is in a simulation which tests its behavior (Nelson, 2007).
- Even unaligned AI will likely model humans in instrumental simulations (Bostrom, 2003a) needed to solve the Fermi paradox.
- Humans could be cost-effective workers in some domains and might be treated as slaves.
- AI could preserve some humans as a potentially valuable asset, perhaps to trade information about them with potential alien AI (Bostrom, 2016a), or to sell them to a benevolent AI.

- AI may still preserve information about human history and DNA for billions of years, even if the AI does not use or simulate humans in the near term. It may later return them to life if it needs humans for some instrumental goal.
- Early AI may use human wetware (biological brains) as an effective supercomputer.
- AI could ignore humans and choose to live in space. Humans would survive on Earth. AI would preserve us if its marginal utility from human atoms is less than its marginal instrumental utility from humans' continued existence.
- As human values are formed by evolution, an evolving AI system (Smart, 2012) may naturally converge to a similar set of values to humans, or basic AI drives (Omohundro, 2008).

Conclusion

There are many possible global solutions to the AI safety problem, but humanity must choose the one that has the highest probability of successful implementation. The current situation of quick development of neural net AIs by large IT companies.

Clearly, a "no AI" solution should not be implemented, as it is immoral and ineffective. As "AI safety theory" is lagging current AI development, controllable, self-improving AI as a global solution will likely not be possible in the next couple of decades. We also lack the global coordination (Bostrom, 2013) to create an AI Nanny, as well as the technologies necessary for human uploading.

Neural-net-based solutions developed by major IT companies are the current are of greatest technological success in AI research (Fernando, 2017), (Shakirov, 2016), ("AlphaGo," 2017). Such organizations not only create AI, but improve their own organizational structure by similar processes, giving rise to "self-improving organizations". Google ("Alphabet") is the leader here by a large margin.

Soft acceleration of several self-improving organizations seems to be the most plausible way to build a mild form of superintelligence in the current epoch, a plan Christiano named "prosaic AI" (Christiano, 2016). It may also be fueled by an AI race between the US and China (Ministry of National Defense of the People's Republic of China, 2016). Thus, the main questions are how to lower the probability of two scenarios: the appearance of a rogue self-improving AI inside a self-improving organization, and AI cold war (and hot war) as result of integration of AI and national defense systems (De Spiegeleire, Maas, & Sweijs, 2017), (Turchin & Denkenberger, 2018b).

Disclaimer

This article represents views of the authors and does not necessarily represent the views of GCRI or ALLFED.

Acknowledgments

We would like to thank Anthony Barrett for insightful comments.

References:

About MIRI. (2017). MIRI. Retrieved from https://intelligence.org/about/

Alexander, S. (2015). Should AI be open. Retrieved from https://slatestarcodex.com/2015/12/17/should-ai-be-open/

Alexander, S. (2016). Ascended economy? Retrieved from http://slatestarcodex.com/2016/05/30/ascended-economy/

AlphaGo. (2017). Deepmind. Retrieved from https://deepmind.com/research/alphago/

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

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

Batin, M., Turchin, A., Markov, S., Zhila, A., & Denkenberger, D. (2018). Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence. *Informatica (Slovenia)*, *41*, 401.

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

Beckstead, N. (2013). *On the overwhelming importance of shaping the far future*. New Brunswick, NJ: Department of Philosophy, Rutgers university.

Berglas, A. (2012). Artificial intelligence will kill our grandchildren (singularity). *Unpublished Manuscript, Draft*, *9*.

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

Bostrom, N. (2003a). Are You Living In a Computer Simulation? *Published in Philosophical Quarterly (2003) Vol. 53, No. 211, Pp. 243-255.*

Bostrom, N. (2003b). Human genetic enhancements: a transhumanist perspective. *The Journal of Value Inquiry*, *37*(4), 493–506.

Bostrom, N. (2006). What is a singleton. *Linguistic and Philosophical Investigations*, *5*(2), 48–54.

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. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31.

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

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

Bostrom, N. (2016b). Strategic Implications of Openness in AI Development. *Working Draft*. Retrieved from http://www.nickbostrom.com/papers/openness.pdf

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

Brin, D. (1998). *The Transparent Society*. Perseus Book.

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

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Filar, B. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv Preprint ArXiv:1802.07228*.

Camarinha-Matos, L. M., & Afsarmanesh, H. (2005). Collaborative networks: a new scientific discipline. *Journal of Intelligent Manufacturing*, *16*(4–5), 439–452.

Christiano, P. (2018, February 24). Takeoff speeds. Retrieved March 5, 2018, from https://sideways-view.com/2018/02/24/takeoff-speeds/

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

Cole, D. D., Denkenberger, D., Griswold, M., Abdelkhaliq, M., & Pearce, J. (2016). Feeding Everyone if Industry is Disabled. In *Proceedings of the 6th International Disaster and Risk Conference. Presented at the 6th International Disaster and Risk Conference*. Davos, Switzerland.

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

Denkenberger, D., Cole, D., Griswold, M., Pearce, J., & Taylor, A. R. (2016). Non Food Needs if Industry is Disabled. In *Proceedings of the 6th International Disaster and Risk Conference. Presented at the 6th International Disaster and Risk Conference*. Davos, Switzerland.

Dewey, D. (2016). Long-term strategies for ending existential risk from fast takeoff. Nov. Retrieved from https://drive.google.com/file/d/1Q4ypVnZspoHTd0OjEJYUHvSvq3O9wjRM/view

Drexler, E., & Phoenix, C. (2004). Safe exponential manufacturing. *Nanotechnology*, *15*, 869.

Fernando, C. (2017). PathNet: Evolution Channels Gradient Descent in Super Neural Networks. *ArXiv:1701.08734 [Cs.NE]*. Retrieved from https://arxiv.org/abs/1701.08734

Future of Life Institute. (2017). Asilomar AI Principles. Future of life institute. Retrieved from https://futureoflife.org/ai-principles/

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.

Hanson, R. (2018). Overcoming Bias : Kaczynski's Collapse Theory. Retrieved February 13, 2018, from http://www.overcomingbias.com/2018/01/kaczynskis-collapse-theory.html

Hanson, R., & Sun, W. (2012). "Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets". *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI2012). ArXiv:1210.4900*.

Hughes, J. (2001). Relinquishment or Regulation: Dealing with Apocalyptic Technological Threats. *Hartford, CT, November*, *14*.

Itut. (2017). Reality Check: 'We Are Not Nearly As Close To Strong AI As Many Believe.' Itutneswlog. Retrieved from http://newslog.itu.int/archives/1566

Jones, S. E. (2013). *Against technology: From the Luddites to neo-Luddism*. Routledge.

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

Kharpal, A. (2017, February 15). Elon Musk: Humans must merge with machines or become irrelevant in AI age. *CNBC*. Retrieved from http://www.cnbc.com/2017/02/13/elon-musk-humans-merge-machines-cyborg-artificial-intelligence-robots.html

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

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

List of semiconductor fabrication plants. (2017). Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_semiconductor_fabrication_plants

Luksha, P. (2014). NeuroWeb Roadmap: Results of Foresight & Call for Action. Slideshare presentation. Retrieved from https://www.slideshare.net/PavelLuksha/neuroweb-roadmap-preliminary

Mahoney, M. (2008). A Proposed Design for Distributed Artificial General Intelligence. Retrieved from http://mattmahoney.net/agi2.html

Marx, K. (1867). *Capital: A Critique of Political Economy. The process of production of capital* (Vol. 1).

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

MIRI. (2016). MIRI AMA - anyone may ask. Retrieved from http://effective-altruism.com/r/main/ea/12r/ask_miri_anything_ama/

Morris, D. Z. (2017). Elon Musk: Artificial Intelligence Is the "Greatest Risk We Face as a Civilization." Retrieved July 18, 2017, from http://fortune.com/2017/07/15/elon-musk-artificial-intelligence-2/

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.*

Nelson, R. (2007). How to Deter a Rogue AI by Using Your First-mover Advantage. SL4. Retrieved from http://www.sl4.org/archive/0708/16600.html.

Nuclearweaponarchive. (2001). India's Nuclear Weapons Program - Smiling Buddha: 1974. Retrieved July 18, 2017, from http://nuclearweaponarchive.org/India/IndiaSmiling.html

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).

Ouagrham-Gormley, S. B. (2013). Dissuading Biological Weapons. In *Proliferation Pages* (pp. 473–500). Retrieved from http://dx.doi.org/10.1080/13523260.2013.842294

Partnership for AI. (2017). Partnership for AI. Retrieved from https://www.partnershiponai.org/

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

Ramamoorthy, A., & Yampolskiy, R. (2018). Beyond MAD?: the race for artificial general intelligence. *ICT Discoveries, Special Issue No. 1.*

Ria Novosti. (2009). Испытания ядерного оружия на Тоцком полигоне. Справка. Retrieved July 18, 2017, from https://ria.ru/defense_safety/20090914/184923659.html

Robots: Legal Affairs Committee calls for EU-wide rules. (2017). European parlaiment. Retrieved from http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules

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

Scott, A. (2016). Book review: Age of Em. Slatesracodex. Retrieved from http://slatestarcodex.com/2016/05/28/book-review-age-of-em/

Shakirov, V. (2016). Review of state-of-the-arts in artificial intelligence with application to AI safety problem. *ArXiv Preprint ArXiv:1605.04232*. Retrieved from https://arxiv.org/abs/1605.04232

Shulman, C. (2011). Arms races and intelligence explosions. In *Singularity Hypotheses*. Springer. Retrieved from

http://singularityhypothesis.blogspot.ru/2011/04/arms-races-and-intelligence-explosions.html

Smart, J. (2012). The transcension hypothesis: Sufficiently advanced civilizations invariably leave our universe, and implications for METI and SETI. *Acta Astronautica Volume 78, September–October 2012, Pages 55–68.* Retrieved from http://www.sciencedirect.com/science/article/pii/S0094576511003304

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

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

Sotala, K. (2017). Disjunctive AI scenarios: Individual or collective takeoff? Retrieved from http://kajsotala.fi/2017/01/disjunctive-ai-scenarios-individual-or-collective-takeoff/

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

Sotala, K., & Yampolskiy, R. (2015). Responses to catastrophic AGI risk: A survey. *Physica Scripta*, *90*(1).

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf.

Teller, E. (1946). LA-602: The ignition of atmosphere with nuclear bombs. Los Alamos National Laboratory. Retrieved from http://large.stanford.edu/courses/2015/ph241/chung1/docs/00329010.pdf

The Telegraph. (2009). Russian spacecraft landed on moon hours before Americans. Retrieved from http://www.telegraph.co.uk:80/science/space/5737854/Russian-spacecraft-landed-on-moon-hours-before-Americans.html

Tomasik, B. (2017). Artificial Intelligence and Its Implications for Future Suffering. Retrieved from https://foundational-research.org/artificial-intelligence-and-its-implications-for-future-suffering

Torres, P. (2016). Problems with defining an existential risk. Retrieved from https://ieet.org/index.php/IEET2/more/torres20150121

Turchin, A. (2017a). *All Hail the King: Human Model Based Global AI Nanny, Which Prevents Creation of the Self-Improving Superintelligence.*

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

Turchin, A., & Denkenberger, D. (2018a). *Classification of Global Catastrophic Risks Connected with Artificial intelligence.* Under review in AI&Society.

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*.

Urban, T. (2017). Neuralink and the Brain's Magical Future. waitbutwhy.com. Retrieved from http://waitbutwhy.com/2017/04/neuralink.html

Venture scanner. (2016). Artificial Intelligence Q1 Update in 15 Visuals. Retrieved from https://venturescannerinsights.wordpress.com/tag/artificial-intelligence-company-list/.

Vinding, M. (2016). *Reflections on Intelligence*. Retrieved from https://www.smashwords.com/books/view/655938

Weng, Y.-H., Chen, C.-H., & Sun, C.-T. (2008). Safety Intelligence and Legal Machine Language: Do We Need the Three Laws of Robotics? In *Service Robot Applications*. InTech.

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

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

Yudkowsky, E. (2001). *Creating Friendly AI 1.0*. Retrieved from intelligence.org/ les/CFAI.pdf

Yudkowsky, E. (2004, May). Coherent Extrapolated Volition. Retrieved from http://intelligence.org/files/CEV.pdf

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. (2010). *Harry Potter and Method of rationality.*

Yudkowsky, E. (2017). There's No Fire Alarm for Artificial General Intelligence. Retrieved January 22, 2018, from https://intelligence.org/2017/10/13/fire-alarm/