big data and cognitive computing

MDPI

Article

Global Solutions vs. Local Solutions for the AI Safety Problem

Alexey Turchin [1,*], David Denkenberger [2] and Brian Patrick Green [3]

[1] Science for Life Extension Foundation, Prospect Mira 124-15, Moscow 129164, Russia
[2] Alliance to Feed the Earth in Disasters (ALLFED), University of Alaska Fairbanks, Fairbanks, AK 99775, USA; ddenkenberger@alaska.edu
[3] Markkula Center for Applied Ethics, Santa Clara University, Santa Clara, CA 95053, USA; bpgreen@scu.edu
* Correspondence: alexeiturchin@gmail.com

Received: 16 December 2018; Accepted: 15 February 2019; Published: 20 February 2019

Abstract: There are two types of artificial general intelligence (AGI) safety solutions: global and local. Most previously suggested solutions are local: they explain how to align or "box" a specific AI (Artificial Intelligence), but do not explain how to prevent the creation of dangerous AI in other places. 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 groups: 1. No AI: AGI 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 "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. Not every local solution scales to a global solution or does it ethically and safely. The choice of the best local solution should include understanding of the ways in which it will be scaled up. Human-AI teams or a superintelligent AI Service as suggested by Drexler may be examples of such ethically scalable local solutions, but the final choice depends on some unknown variables such as the speed of AI progress.

Keywords: AI safety; existential risk; AI alignment; superintelligence; AI arms race

1. Introduction

The problem of how to prevent a global catastrophe associated with the expected development of AI of above human-level intelligence is often characterized as "AI safety" [1]. The topic has been explored by many researchers [2–5]. Other forms of "AI safety," typically associated with narrow-AI such as that for self-driving cars or other narrow applications, are not considered in this paper.

An extensive review of possible AI safety solutions has been conducted by Sotala and Yampolskiy [4]. In their article, they explore a classification of AI safety solutions by 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, for example, global technological relinquishment or use of the first superintelligent AI to prevent other AIs from arising. Most solutions described by Sotala and Yampolskiy [4] are considered local solutions in our classification scheme.

Recent significant contributions to the global solutions problem include Christiano's slow takeoff model [6], which demonstrated that such a takeoff could happen earlier than a fast takeoff; Ramamoorthy and Yampolskiy's research on AI arms races: "Beyond MAD?: the race for artificial general intelligence" [7]; Brundage et al.'s "The Malicious Use of Artificial Intelligence" [8]; and research on a collective takeoff by Sotala [9]. The problem of "other AIs," central to the global AI safety conundrum, has been explored by Dewey [10], who suggested four types of solution: international coordination, sovereign AI (superintelligent AI acting independently on global scale), an AI-empowered project, and some other decisive technological advantage.

Any local safety solution which cannot be applied globally cannot in itself determine the course of human history, as many other AIs may 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 the only global power: *Singleton* [11].

Generally, when we use the term "AI" throughout this article, we do not mean standard contemporary systems of machine learning processing Big Data, but rather the descendants of these contemporary systems, which are dramatically more sophisticated, nuanced, and process vastly more data even faster, and therefore attain intelligence equivalent to and/or surpassing human intelligence. Additionally, this article is based on the assumption—shared by many (e.g., [3,12]), but not all AI researchers—that above human-level AI is possible in the relatively near future (21st century) and the world's socio-political structure will be approximately the same as now at the moment of its creation. This assumption about the possibility of superhuman AI is naturally followed by concerns about the safety of such systems, which may generate not only isolated accidents, but a full variety of possible global catastrophes as explored in Reference [13].

The main thesis of this article is that there are two main types of AI safety solutions: local and global, and that not every local solution scales to a global solution or does it ethically and safely. The choice of the best local solution should include an understanding of the ways in which it may be scaled up. Human-AI teams or a superintelligent AI Service as suggested by Drexler [14] may be examples of such ethically scalable local solutions, but the final choice depends on some unknown variables such as the speed of AI progress [15].

To solve the problem of the relation between global and local solutions, we created a classification of global solutions, which is a simpler task as all global solutions depend on the one main variable: how many different AI systems will be eventually created. We used this classification to identify pairs of local and global solutions, which are less risky when combined.

In Section 2 we overview various levels of AI safety. In Section 3 we look at solutions involving the prevention of AI, 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 technologically modified human beings or human-mind models collaborate directly with or control AI, "inside" it.

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 [16] 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 [17], but low realization is not the same as extinction [18], and Green argues that even full human flourishing is not enough to eliminate existential risk unless ethical standards and practices are also somehow concomitantly perfected [19].

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 [20],

but centered on suffering. Our classification is based on levels of human well-being, the first and most basic of which is survival:

1. "Last man": At least one human being survives creation of strong AI, for example, possibly as an upload, or in a "zoo".
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 [13] or leaves Earth.
3. "No AI": Any outcome where global catastrophe connected with AI has not occurred because there is no AI to provoke the catastrophe. This is a world in which a comprehensive ban on AI is enforced, or AI technologies otherwise never progress to AGI or superintelligence.
4. "Better now": Human civilization is preserved after AI creation in almost the same form in which it exists now, and benefits from AI in many ways, including through the curing of diseases, slowing aging, preventing crime, increasing material goods, achieving interstellar travel, etc. Outcomes in this category likely involve a type of "AI Nanny" [21].
5. "Infinite good": Superintelligent AI which maximizes human values (Benevolent AI) will reach the maximum possible positive utility for humans, but contemporary humans cannot now describe this utility as it is beyond our ability to imagine, as presented by Yudkowsky [22].

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 Levels 1 and 2 are unacceptable as they produce unimaginable human suffering and risk human extinction.

3. "No AI" Solutions

In our world of quick AI development, AI relinquishment seems improbable or requires some unethical and/or risky acts of Luddism. Many of these solutions have been explored by Sotala amd Yampolskiy [4].

Overview of restrictive solutions where advance AI creation is prevented globally:

- International ban
- Legal relinquishment
- Technical relinquishment or AI appears to be not technically possible
- Destruction of capability to produce AI anywhere in the world
    - War
    - Luddism
    - Staging small catastrophe
- Slowdown of AI creation
    - Economical
    - Technology slowdown
    - Overregulation
    - Brain drain from the field
    - Defamation of idea of AI, AI winter

3.1. Legal Solutions, Including Bans

Not many argue for a global AI ban as it is unfeasible under current conditions [23] and would likely only help bad actors [24]. One could imagine that global legal regulation could ban the creation of self-improving agents. However, 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 a group of the most powerful countries. However, 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 [13]. However, there is currently no agreement as to what factors would serve as a credible "alarm," and such agreement may be impossible [25].

Some governmental and non-governmental groups are working to develop guidelines in this area. The EU is considering legislation about robotic ethics [26]. 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 [27] 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 [28]. 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 similar to Open AI, as described by Reference [29], but enforced by the UN.

To implement such an AI agency, the UN would need a powerful enforcement agency. In the same way as when the IAEA fails, an international coalition would need to be able to use sanctions (as against Iran and North Korea) or military intervention, as 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 must be empowered to use force, possibly including cyber weapons or possibly even nuclear weapons. However, in the current world climate, there will be little or no support for the creation of a world government authorized to use powerful weapons to destroy AI labs based only on theory. The only chance for its creation might be if some spectacular AI accident happened, for example, if a narrow-AI-based virus with machine learning capabilities hacked hundreds of airplanes and crashed them into nuclear power plants. In such a case, a global ban on advanced 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 [30–32]. 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 access
- electricity

The rarest commodity are chip fabs, which cost billions of dollars and are needed to create new processors. There are around 200 chip fabs in the world now [33]. If they were closed, no new computers could appear in the world, which might drastically slow AI progress. However, the effect of fabs is rather indirect, as it is possible that enough computers already exist to create AI, especially given the large existing supply of graphics cards, but these chips are simply not in the right configuration.

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 a network) are probably unable to support AI, so if powerful computers and internet connections are switched off, it could considerably slow down AI creation. These restrictions, such as those discussed in Section 3.1,

would require preexisting global coordination. In addition, they would obviously have significant economic consequences.

As a last point, in this section it is worth noting that full AI may not be technically possible in any realistic sense, and if that is the case then whatever "commodity" permits its creation is restricted in a complete sense.

3.3. Destructive 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 (a conventional attack could also be attempted, but would be slower and have a lower probability of success). 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. Nuclear attack of this type is extremely unlikely, unless it were perceived that an "AI uprising" had already started.

Similar to the first option, but with a more purely anti-electronics approach, 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, fossil fuel extraction, and 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 [34,35]. However, recovery of technological civilization and thus the ability to recreate AI is possible, 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, even if it succeeded, would likely be temporary.

One could imagine other means of destruction, ranging from economic recession to Luddism [36], to various global catastrophes, but, as with the above options, all of them are impractical and morally unacceptable. In the future, perhaps some high-tech methods of AI halting might be implemented, such as the Stuxnet computer virus that destroyed Iran's uranium centrifuges [37]. A virus could be used to destroy chip fabs, shut down the internet, or cut electricity. There are other ideas in the field, but an exhaustive list is not within the scope of this paper.

As one last point, the unilateralist's curse—the lack of coordination between many actors with the same goal [38]—may exaggerate activities of those groups that at least believe in the possibility of safe AI.

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, include:

- Public fears of AI.
- The next AI winter, lack of interest in its development (there have already been two after hype in the 1960s and 1980s).
- Extensive regulation of the field.
- Intentional disruption of 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," where the 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 as though the time of marginalization of AI has passed.

- Lastly, depending on the technical challenges, advanced AI, including AGI and superintelligence, may not be technically possible in the near future, although there is no reason at this point to assume it is not. However, if these challenges appear, AI could be indefinitely delayed.

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 due to the global power of the singleton.

Overview of "one AI" solutions:

- First AI is used to take over the world
    - First AI is used as a military instrument
    - First AI gains global power via peaceful means
        - Commercial success
        - Superhuman negotiating abilities
    - Strategic advantage achieved by narrow AIs produces global unification, before the rise of superintelligent AI, by leveraging preexisting advantage of a nuclear power and increasing first-strike capability
    - First AI is created by a superpower and provides it a decisive strategic advantage
    - First AI is reactive, and while it does not 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 an 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)
    - Comprehensive AI Services
    - Distributed AI based on blockchain (SingularityNET)
    - AI as technology everywhere (openness)
    - Augmented humans as AI neurons (Neuralink)
    - Superintelligence as a distributed optimization process by rivalry 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

4.1. First AI Seizes World Power

Advanced agential AIs will be able to act in the world autonomously. Superintelligent AI could potentially seize 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 [39].

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 [40].

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 negotiating ability as a service, which everyone will be interested in applying, and in that case, there will be no need for a military world takeover.

4.1.1. Concentrate the Best AI Researchers to Create a Powerful and Safe AI First

This idea is to create something similar to 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. For example, scientists involved in the project created the LA-602 report about the possibility of causing a nuclear-initiated chain reaction in the atmosphere [41].

Later efforts to create nuclear weapons in other countries were not so safety-oriented. The Soviets exploded a bomb over their own troops [42]. The Indians dropped explosives intended to be part of their first nuclear bomb during critical assembly—fortunately, it did not detonate [43].

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 organization, but AI is created by another group.

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 other words, the leader would have more time to think about safety, by virtue of their being in the lead.

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 [44,45] to be implemented elsewhere.

4.1.2. Using the Decisive Advantage of Non-Self-Improving AI to Create an AI Nanny

Sotala [46] wrote that even non-self-improving AI may gain a decisive strategic advantage if it is effective at designing new weapons, or in strategic military or 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 might be built around a human upload or its equivalent, which gains most of its power not from self-improvement, but from running on high-speed hardware. Such a high-speed human analogue gaining global power via social manipulation and designing new weapons might become an "AI king".

One way to gain such a decisive strategic advantage would be if the first AI were created by a superpower (either China or the 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 (e.g., access to secret information, control of nuclear weapons, large financial resources). Exemplifying this view, see the recent remark by Vladimir Putin that "the nation that leads in AI 'will be the ruler of the world'" [47].

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 a way that yields a high probability of no retaliation.

Dewey [10] suggested the first AI could be reactive or proactive: Proactive AI prevents creation of other AIs, starting preemptive wars against them, and reactive AI only limits or ensures the safety of other AI fast takeoffs. 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.

4.1.3. Risks of Creating Hard-takeoff AI as a Global Solution

In AI safety research, it is often 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 then prevent the creation of another AI in, say, China. From the point of view of international law, such an action by an AI could be an act of war [40].

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 [48] 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 an AI could be classified as a cyberweapon) and may prevent open dissemination of any AI safety theory based on such a suggestion.

It appears that creation of a military infrastructure is a convergent instrumental goal for any first AI [40]. This infrastructure would help the AI prevent the 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. Therefore, 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 below.

4.2. One Global AI Created by Collective Efforts

4.2.1. 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'" [21]. 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 [49] 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, and the forceful integration of any remaining "rogue states." This could resemble the current dynamic between a large international coalition of nuclear-armed states 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 might appear if one country gained 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 nanotechnological 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 appeared, but spectacularly failed 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 new data 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 [50] as a very large and prohibitively expensive international project, for example, 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 [51], in a sense similar to Baruch's 1946 plan to centralize nuclear research [10].

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. Therefore, an AI Nanny would likely be rather safe and under international control. However, the opportunity for such a project may be 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. Yet the potential is not completely lost; large international collaborations such as the "Partnership on AI" [52] could contribute momentum to the creation of an AI Nanny, if they chose to do so.

4.2.2. Levels of Implementation of the AI Nanny Concept

We suggest four 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 could be to use an upload human as an "AI king," or world governor, with the main mission of preventing the creation of other AIs [53]. Such an AI king would run at higher speeds than ordinary humans, using all available hardware, which will give it greater intelligence while maintaining alignment with human values. This idea would be obviously 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 a high-intelligence AI Nanny as described by Reference [21]. This AI would be some form of superintelligence (SI), as much above humans as humans are above apes. In this case,

there would be exactly the same problems as with the control of any other strong AI [49]. However, if the system were weaker, it might be possible to find Goldilocks' path between its ability to control research and our ability to control the system.

4.2.3. Global Transition into AI: Non-Agential AI-Medium Everywhere, Accelerating Smoothly without Tipping Points

The AI described above was agential. However, 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 a 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 [54]. This feature could still be devastating, as we know that in an analogous case, market forces can destroy traditional cultures more effectively than weapons [55]. 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 [56], and the network of scientific references [57]. One of the routes to an AI-medium could be to connect all human brains through some form of network, producing, in effect, a global brain [58].

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, for example, Sotala's review criticizing Vinding's recent book discussing the difference between individual and collective takeover [9,59].

Scott Alexander argues that an accelerating self-improving AI-medium is possibly a negative outcome as it could take the form of an "ascending economy" [55], where a group of market agents create an evolving ecosystem, which destroys all human values in order to increase "growth." Karl Marx criticized market economics for the same flaw [60].

As the AI-medium naturally evolves without taking into account human values, it cannot be considered friendly or unfriendly to humans. Given this, unless regulated in some way, it will either prioritize human survival, if humans will be able to positively interact with it, or it will ignore humans. John Smart [61] predicted that the evolution of such a system will consist of constant acceleration and miniaturization, which could be described by a hyperbolic law.

Drexler suggested another form of AI-medium, Comprehensive AI Services [14], in which superintelligence does not have agency. Instead, it consists of many narrow superintelligent tools, which also could be used to create needed level of surveillance to prevent rogue AI appearance elsewhere. A primitive example of such service now is Google with its many "Tool AIs": web search, email, drive, which are integrated in one ecosystem but are not agential. However, as Gwern wrote [62], any Tool AI "wants" to be agential AI, as it would increase its efficiency, and thus AI Services could eventually turn into or spawn potentially dangerous agential AI.

4.3. Help Others to Create Safe AI

4.3.1. Promoting Ideas of AI Safety in General and the Best AI Safety Solution to All Players

Helping others develop improved AI safety is a global solution if there are ways to reach all significant AI players.

As we mentioned above, it is a priori improbable that the same team that creates an optimal AI safety theory will also create the first AI unless it is part of an international collaboration. Therefore, teams working on AI safety should try to convince other AI teams to adopt the best AI safety theory.

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

- Funding of AI safety research.
- Promotion of the idea of AI safety.
- Protesting military AI.
- Friendly AI training for AI researchers.
- Providing publicly available safety recommendations.
- Increasing the "sanity waterline" and rationality in the general population and among AI researchers and policymakers.
- Lowering global levels of confrontation and enmity.
- Forming political parties for the prevention of existential risks and control of AI risks, or lobbying current political parties to adopt these positions. However, even if such parties were to win in larger countries and were able to change policy, there would still be countries that could use any technology "freeze" in larger countries to their 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 'Harry Potter and the Methods of Rationality' [63] 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 quantity and quality of thought working on AI safety theory
- Establishing relationships between the best AI teams, as some of the people who will have worked on AI safety may have come from such teams, may eventually join them, or may otherwise have friends there, and
- Promoting the idea that unlimited self-improvement is dangerous and unstable for all players, including AIs.

4.3.2. Selling AI Safety Theory as an Effective Tool to Align Arbitrary AI

One possible way to reach many people is to make the solution attractive. If AI safety implementation can be used to 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 place goals into the AI. The AI creator could save many resources by implementing a proven alignment method. However, while this lessens the probability that the AI will run amok, the creator could still 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 support could increase its implementability.

4.4. Local Action to Affect Other AIs Globally

4.4.1. Slowing the Appearance of Other AIs

In this scenario people could take 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.

4.4.2. Ways to Affect a Race to Create the First AI

An AI creation race is generally regarded as bad because it encourages the creation of the least-safe AIs first. A war between AIs may also become possible if several AIs are created simultaneously [64,65].

There are many ideas on how to affect an AI race in order 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, for example, openness. Bostrom has shown that if no one knows the capabilities of others and their own capabilities, it will slow down the race, so openness about capabilities may be dangerous [66].

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

- Changing the number of participants.
- Increasing or decreasing information exchange and level of openness.
- Reducing the level of enmity between organizations and countries, and preventing conventional arms races and military buildups.
- Increasing the level of cooperation, coordination, and acceptance of the idea of AI safety among AI researchers.
- Changing the total amount of funding available.
- Promoting intrinsic motivations 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" [67]; 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.
- Changing social attitudes toward the problem and increasing awareness of the idea of AI safety.
- 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.
- Affecting the idea of the AI race as it is understood by the participants [67]: 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 might become 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 [68].
- Affecting the public image of AI researchers who are currently presented as not wanting beneficial AI design [67].
- Refraining from suggestions of draconian surveillance as they "inadvertently frame efforts to promote beneficial AI as being the problem, not the solution" [67].
- Stigmatization of building recursive self-improving AI by framing them as morally unacceptable, as has been done with landmines. The stigma impelled even countries that did not sign the treaty that prohibits landmines to reduce production [67].

- Deliberate association with crackpottery: an example is UFO (Unidentified Flying Objects) research: anyone who mentions the word "UFO" will no longer be accepted in the scientific community as a credible scientist. This partially worked against AI during past AI winters, when scientists tried not to mention the words "artificial intelligence." Society could come to associate "self-improving AI" with craziness, which would be not difficult if we pick some of the most outstanding ideas from associated internet forums, e.g., Roko's Basilisk [69]. 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 in neural nets by DeepMind show that the idea of self-improving AI is becoming mainstream [70].
- Affecting 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.

4.4.3. Participating in Acausal Deals with Future AI

Rolf Nelson [71] 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 to display benevolence. However, such an approach would probably work only for an AI singleton, and it is our last level of defense.

5. "Many AI" Solutions

5.1. Overview of the "Net Solutions" of AI Safety

5.1.1. 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 become obsolete in the age of AI [72]. However, there are risks to this approach [66].

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

1. The combined intelligence of many AIs (the net of AIs) is much higher than the one of any rogue AI, so the net is able to create effective protection. An AI-net could form something similar to AI "police," which prevent any single AI from unlimited growth. This is analogous to the way the human body provides a multilevel defense against unlimited growth of a single cancerous cell in the form of an immune system. The approach is somewhat similar to the AI Nanny approach [21], but an AI Nanny is a single AI entity. An AI-net consists of many AIs, which use ubiquitous transparency [73] to control and balance [74] each other.
2. Value diversity among many AI-sovereigns [2,75] guarantees that different positive values will not be lost. Different members of the net have different terminal values, thus ensuring diversity of values, as long as the values do not destructively interfere. If the values do destructively interfere, then solutions must be found for these conflicts.

We will call this many AI solution a "Multipolar Singleton" [11], 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.
- Because Earth is surrounded by infinite space, different AIs could start to travel to the stars in different directions, and as each direction includes a very large number of stars, even very ambitious goals could be not mutually exclusive and might not provoke conflicts and wars.
- Finding it mutually beneficial to create 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 reduce to a single AI dictatorship.

5.1.2. The Importance of Number in the 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 creators, as well as the upper limit of individual intelligence, if it exists. 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:

- Two AI-sovereigns' semi-stable solution, similar to the Cold War [76].
- From several to dozens of sovereign AIs, similar to existing nation-states; they may be evolved from nation-states, or from large companies.
- From thousands to billions of AIs, with relations similar to relations between humans now, possibly resulting from some brain uploading technology [74], human augmentation [77], or genetic modification [78], but each single AI is not significantly above human level.
- Uncountable or almost-infinite number of AIs, similar to AI-medium, discussed above. This could be similar to the IoT, but with AIs as nodes.

In the following list we present an overview of possible solutions, which will be explored in detail below.

- Net of AIs forms a multilevel immune system to protect against rogue AIs and has a diversity of values, thus including human-positive values
    - Instruments to increase the 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
    - Net of AIs is based on human uploaded minds

- Several AI-sovereigns coexist, and they have better defensive than offensive capabilities
    - Two AIs semi-stable "Cold war" solution, characterized by
        - tight arms race
        - military AI evolution
        - MAD defense posture
    - AI-sovereigns appear from nation-states
        - very slow takeoff and integration with governments
    - Different AIs expand in space in different directions without conflict
    - Creation of AIs on remote planets

5.2. From the Arms Race between AI-Creating Teams to the Net of AIs

5.2.1. 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" [29]. 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 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 that 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 the benefits of the membership in the net, namely access to the results and capabilities of others. However, this depends on how much information becomes part of the public domain. The open net will have an "intelligence advantage" over any smaller player, which makes it more probable that self-improvement will start inside the open net, or that the net will have time to react before a 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 assess and possibly control all the low-hanging fruits of self-improvement, for example, 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 stopped publishing articles about uranium in 1939 showed that they were trying to keep their work secret and therefore indirectly 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 [73]. 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 [66]. It would also not be easy to balance a dangerous AI, as it could undertake local actions that could quickly kill everybody, like constructing a very large nuclear cobalt bomb [79] or a dangerous biological virus. However, 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 [80].

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

5.2.2. 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 that it is better if they do not know about each other's or their own capabilities [64].

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 three hours of each other on 14 February 1876 [81] and the Soviet–US race to bring material back to Earth from the Moon was decided by three days in 1969 [82].

Increasing the number of independent AI projects will increase the probability that several of them will have hard takeoffs simultaneously, but it will also increase the chances that some of the programs will have a very low level of safety, as Bostrom et al. note [64].

The publicity around AI in recent years has likely contributed to the growth of AI companies. Venture Scanner tracked 957 AI-creating companies [83]. While most of them are not trying to build AGI, many of them would be happy to have an 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 on the list are fake or should not be counted for other reasons. Therefore, 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.

5.2.3. Change of the Self-Improving Curve Form, So That the Distance Between Self-Improving AIs Will Diminish

Yampolskiy has argued that there are several reasons why the actual self-improving of one AI system may be described by a logarithmic rather than exponential curve [84]. However, artificial interventions such as taking low-hanging fruits or 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.

5.3. Instruments to Make the Net of AIs Safer

5.3.1. 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, such as a nanofab, which is regarded a safer form of nanotech than nanorobots [85].

One possible design of such a "robotic brain" could be a human upload [74] or some simplified model of a human brain, which finds a balance between upload and neuromorphic AI [53].

5.3.2. Starting Many AIs Simultaneously

Any AI-creating team could start not one, but many AIs, just to balance the possible flows in the first AI or to observe its possible flaws depending on initial conditions of different AI. Such an approach will likely have unpredictable consequences, and might 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 with slight changes of goal function. However, the idea of "beneficial computer viruses" was already discussed and it was concluded that such viruses would not be better than normal antivirus software, as the second virus, intended to deactivate the first one, will spread much less and will also cause harm [86].

6. Solutions in Which Humans are Part of the AI System

6.1. Different Ways to Incorporate Humans inside AI

Some form of superintelligence may be created with humans as participants within it, as Drexler suggested in his Comprehensive AI Services. However, as Bostrom shows [2], there is always the problem of the "second transition," that is, the appearance of a more powerful AI inside such a system, one which no longer needs humans. So, any such system would need to create an AI police to prevent a "second transition."

Another problem is that most such solutions are lagging, as human uploading is technically still far away, if possible at all.

Some ways of incorporating humans inside AI include:

- AI could be built around a human core or as a human emulation. It could result from effective personal self-improvement via neural implants [2], adding tool AIs and exocortex. There is no problem of "AI alignment," as there are not two agents that should be aligned, but only one agent whose value system is evolving [53], however, if the human core is not aligned with the rest of humanity, the same misalignment problem could appear—therefore the ethics of the core human are crucial.
- AI could appear from a net of self-improving posthumans, connected via neural interfaces [87]. This combines ideas of social networks, blockchain, and Neuralink [77]. Such a net could conceivably appear from the evolution of some types of medical AI [88].
- AI could result from genetic modification of humans for intelligence improvement [78].
- 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 [74,89].
- Only one human upload is created, and it works as an AI Nanny, preventing the emergence of any other superintelligences [53].
- 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 [29] or the "Partnership on AI".

- 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, or a multipolar world. In this scenario, we could 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 at this point, but there could also be scenarios which combine some of the ideas in this section.

6.2. 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 [71].
- Even unaligned AI will probably model humans in instrumental simulations [90] needed to solve the Fermi paradox.
- Humans could be cost-effective workers in some domains and might therefore be retained, though only to be treated as slaves.
- AI could preserve some humans as a potentially valuable asset, perhaps to trade information about them with potential alien AI [75], 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.
- AI may use human "wetware" (biological brains) as efficient supercomputers.
- AI could ignore humans and choose to live in space, while humans would survive on Earth. AI would preserve humanity if the marginal utility derivable from humanity's atoms is less than the marginal instrumental utility from humanity's continued existence.

As human values are formed by evolution, an evolving AI system [61] may naturally converge to a similar set of values as humans [91].

7. Which Local Solutions Are the Best to Get a Stable Global Solution?

In the sections above, we overviewed all (to the best of our knowledge) previously suggested global solutions for AI safety.

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.

Clearly, a "no AI" solution should not be implemented, as it would be unethical (due to opportunity costs) and ineffective (due to intense pressures to achieve AI). As "AI safety theory" is lagging current AI development; a controllable, self-improving AI as a global solution will probably not be possible in the next couple of decades. We also lack the global coordination [92] to create an AI Nanny, as well as the technologies necessary for human uploading.

Neural-net-based solutions developed by major IT companies are currently the greatest technological source of success in AI research [70,93,94]. 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" [50]. It may also be fueled by an AI race between the US and China [95].

In the current technological and political situation, several local approaches seem to be most safely scalable to the global scale:

(1) Comprehensive AI Services, which could become a basis for a system of ubiquitous surveillance and AI Police, preventing appearance of rogue AIs.
(2) Research in human uploads or human-mind models, which will result in many AIs of relatively limited capabilities [74]. This again could be used to create AI Police.
(3) Self-improving organizations, where humans and AI work together which is basically is part of Drexler's suggestion [14], but also could be done in Christiano's approach of iterated amplification and factored cognition [96].
(4) Robotic mind-bricks, that are pre-trained AI with limited capabilities and prefabricated safety measures which would be sold widely and provide a basis for global AI policing.
(5) AI Safety as a service, similar in some sense to current antivirus computer industry.

8. Conclusions

Suggested solutions to AI safety problem are either local or global, and when choosing a local solution, we also should take into account how it could be safely scaled globally. In this article, we posed the problem of relation between global and local solutions, overviewed existing global solutions and estimated their safety.

We identified a group of local solutions which seems to be more easily and safely scaled into a global level. This group includes such approaches as Comprehensive AI Services, selling robotic mind-bricks, and AI Safety as a service.

9. Disclaimer

This article represents views of the authors and does not necessarily represent the views of the ALLFED, the Markkula Center for Applied Ethics, or other organizations to which the authors belong.

Author Contributions: A.T. provided writing—original draft preparation, D.D. provided writing—review and editing, and B.P.G. provided writing—review and editing.

Funding: This research received no external funding.

Acknowledgments: We thank Anthony Barrett for insightful comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yampolsky, R.; Fox, J. Safety engineering for artificial general intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
2. Bostrom, N. *Superintelligence*; Oxford University Press: Oxford, UK, 2014.
3. Russell, S. 3 Principles for Creating Safer AI. Available online: https://www.youtube.com/watch?v=EBK-a94IFHY (accessed on 18 February 2019).
4. Sotala, K.; Yampolskiy, R. Responses to catastrophic AGI risk: A survey. *Phys. Scr.* **2015**, *90*, 069501. [CrossRef]
5. Yudkowsky, E. *Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks*; Cirkovic, M.M., Bostrom, N., Eds.; Oxford University Press: Oxford, UK, 2008.
6. Christiano, P. Takeoff Speeds. Available online: https://sideways-view.com/2018/02/24/takeoff-speeds/ (accessed on 5 March 2018).
7. Ramamoorthy, A.; Yampolskiy, R. Beyond MAD?: The race for artificial general intelligence. *ICT Discov. Spec. Issue* **2018**, *1*, 1–8.

8. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv* **2018**, arXiv:1802.07228.
9. Sotala, K. Disjunctive AI Scenarios: Individual or Collective Takeoff? 2017. Available online: https://kajsotala.fi/2017/01/disjunctive-ai-scenarios-individual-or-collective-takeoff/ (accessed on 18 February 2019).
10. Dewey, D. *Long-Term Strategies for Ending Existential Risk from Fast Takeoff*; Taylor & Francis: New York, NY, USA, 2016.
11. Bostrom, N. What is a singleton. *Linguist. Philos. Investig.* **2006**, *5*, 48–54.
12. Krakovna, V. Risks from general artificial intelligence without an intelligence explosion. *Deep Saf.* **2015**, *26*, 1–8.
13. Turchin, A.; Denkenberger, D. Classification of Global Catastrophic Risks Connected with Artificial intelligence. *J. Br. Interpanet. Soc.* **2018**, *71*, 71–79. [CrossRef]
14. Drexler, K.E. Reframing Superintelligence. Available online: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf (accessed on 18 February 2019).
15. Turchin, A. Assessing the future plausibility of catastrophically dangerous AI. *Futures* **2018**. [CrossRef]
16. Beckstead, N. *On the Overwhelming Importance of Shaping the Far Future*; Department of Philosophy, Rutgers University: New Brunswick, NJ, USA, 2013.
17. Bostrom, N. Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. *J. Evol. Technol.* **2002**, *9*, 2002.
18. Torres, P. Problems with Defining an Existential Risk. Available online: https://ieet.org/index.php/IEET2/more/torres20150121 (accessed on 18 February 2019).
19. Green, B.P. The Technology of Holiness: A Response to Hava Tirosh-Samuelson. *Theol. Sci.* **2018**, *16*, 223–228. [CrossRef]
20. Tomasik, B. *Artificial Intelligence and Its Implications for Future Suffering*; Foundational Research Institute: Basel, Switzerland, 2017.
21. Goertzel, B. Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? *J. Conscious. Stud.* **2012**, *19*, 96–111.
22. Yudkowsky, E. Coherent Extrapolated Volition. Available online: http://intelligence.org/files/CEV.pdf (accessed on 18 February 2019).
23. Weng, Y.-H.; Chen, C.-H.; Sun, C.-T. Safety Intelligence and Legal Machine Language: Do We Need the Three Laws of Robotics. In *Service Robot Applications*; InTech: Rijeka, Croatia, 2008.
24. Hughes, J. Relinquishment or Regulation: Dealing with Apocalyptic Technological Threats. *Hartford CT Novemb.* **2001**, *14*, 06106.
25. Yudkowsky, E. *There's No Fire Alarm for Artificial General Intelligence*; Machine Intelligence Research Institute: Berkeley, CA, USA, 2017.
26. Robots: Legal Affairs Committee Calls for EU-Wide Rules. Available online: http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules (accessed on 18 February 2019).
27. Future of Life Institute Asilomar AI Principles. Available online: https://futureoflife.org/ai-principles/ (accessed on 18 February 2019).
28. Morris, D.Z. Elon Musk: Artificial Intelligence Is the "Greatest Risk We Face as a Civilization". Available online: http://fortune.com/2017/07/15/elon-musk-artificial-intelligence-2/ (accessed on 18 July 2017).
29. Brockman, G.; Sutskever, I. Introducing OpenAI. Available online: https://openai.com/blog/introducing-openai/ (accessed on 18 February 2019).
30. Berglas, A. Artificial intelligence will kill our grandchildren (singularity). Unpublished work, 2012.
31. Green, B. Are science, technology, and engineering now the most important subjects for ethics? Our need to respond. In Proceedings of the 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, Chicago, IL, USA, 23–24 May 2014; pp. 1–7.
32. Green, B. Emerging technologies, catastrophic risks, and ethics: three strategies for reducing risk. In Proceedings of the 2016 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS), Vancouver, BC, Canada, 13–14 May 2016.
33. List of Semiconductor Fabrication Plants. Available online: https://en.wikipedia.org/wiki/List_of_semiconductor_fabrication_plants (accessed on 18 February 2019).

34. Cole, D.D.; Denkenberger, D.; Griswold, M.; Abdelkhaliq, M.; Pearce, J. Feeding Everyone if Industry is Disabled. In Proceedings of the 6th International Disaster and Risk Conference, Davos, Switzerland, 28 August–1 September 2016.
35. Denkenberger, D.; Cole, D.; Griswold, M.; Pearce, J.; Taylor, A.R. Non Food Needs if Industry is Disabled. In Proceedings of the Proceedings of the 6th International Disaster and Risk Conference, Davos, Switzerland, 28 August–1 September 2016.
36. Jones, S.E. *Against Technology: From the Luddites to Neo-Luddism*; Routledge: Abingdon, UK, 2013; ISBN 1-135-52239-1.
37. Kushner, D. The real story of stuxnet. *IEEE Spectr.* **2013**, *50*, 48–53. [CrossRef]
38. Bostrom, N. The Unilateralist's Curse: The Case for a Principle of Conformity. Available online: http://www.nickbostrom.com/papers/unilateralist.pdf (accessed on 18 February 2019).
39. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf: New York, NY, USA, 2017.
40. Turchin, A.; Denkenberger, D. Military AI as convergent goal of the self-improving AI. In *Artificial Intelligence Safety and Security*; CRC Press: Louisville, KY, USA, 2018.
41. Teller, E. *LA-602: The Ignition of Atmosphere with Nuclear Bombs*; Los Alamos Laboratory: Los Alamos, NM, USA, 1946.
42. Ria Novosti Испытания ядерного оружия на Тоцком полигоне. Справка. Available online: https://ria.ru/defense_safety/20090914/184923659.html (accessed on 18 July 2017).
43. Nuclearweaponarchive India's Nuclear Weapons Program—Smiling Buddha: 1974. Available online: http://nuclearweaponarchive.org/India/IndiaSmiling.html (accessed on 18 July 2017).
44. MIRI. MIRI AMA—Anyone May Ask. Available online: http://effective-altruism.com/r/main/ea/12r/ask_miri_anything_ama/ (accessed on 20 February 2019).
45. MIRI. About MIRI. Available online: https://intelligence.org/about/ (accessed on 18 February 2019).
46. Sotala, K. Decisive Strategic Advantage without a Hard Takeoff. 2016. Available online: https://kajsotala.fi/2016/04/decisive-strategic-advantage-without-a-hard-takeoff/ (accessed on 18 February 2019).
47. Putin, V. Open Lesson "Russia Looking to the Future". Available online: http://kremlin.ru/events/president/news/55493 (accessed on 28 October 2017).
48. Kahn, H. *On Thermonuclear War*; Princeton University Press: Princeton, NJ, USA, 1959.
49. Muehlhauser, L.; Salamon, A. Intelligence Explosion: Evidence and Import. In *Singularity Hypotheses*; Springer: Berlin/Heidelberg, Germany, 2012.
50. Christiano, P. Prosaic AI Alignment. Available online: https://ai-alignment.com/prosaic-ai-control-b959644d79c2 (accessed on 18 February 2019).
51. Itut Reality Check: 'We Are Not Nearly As Close To Strong AI As Many Believe'. Available online: https://news.itu.int/reality-check-not-nearly-close-strong-ai-many-believe/ (accessed on 18 February 2019).
52. Partnership for AI. Available online: https://www.partnershiponai.org/ (accessed on 18 February 2019).
53. Turchin, A. Human Upload as AI Nanny 2017. Available online: https://www.academia.edu/38386976/Human_upload_as_AI_Nanny (accessed on 19 February 2019).
54. Mahoney, M. A Proposed Design for Distributed Artificial General Intelligence. 2008. Available online: http://mattmahoney.net/agi2.html (accessed on 18 February 2019).
55. Alexander, S. Ascended Economy? Available online: http://slatestarcodex.com/2016/05/30/ascended-economy/ (accessed on 18 February 2019).
56. Hanson, R.; Sun, W. Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets. *arXiv* **2012**, arXiv:1210.4900.
57. Camarinha-Matos, L.M.; Afsarmanesh, H. Collaborative networks: a new scientific discipline. *J. Intell. Manuf.* **2005**, *16*, 439–452. [CrossRef]
58. Luksha, P. NeuroWeb Roadmap: Results of Foresight & Call for Action. 2014. Available online: https://dlib.si/details/URN:NBN:SI:DOC-IXKS9ZQW (accessed on 18 February 2019).
59. Vinding, M. *Reflections on Intelligence*; Heinemann: Portsmouth, NJ, USA, 2016.
60. Marx, K. Capital: A Critique of Political Economy. The Process of Production of Capital. 1867. Available online: https://oll.libertyfund.org/titles/marx-capital-a-critique-of-political-economy-volume-i-the-process-of-capitalist-production (accessed on 18 February 2019).
61. Smart, J. The transcension hypothesis: Sufficiently advanced civilizations invariably leave our universe, and implications for METI and SETI. *Acta Astronaut.* **2012**, *78*, 55–68. [CrossRef]

62. Gwern. Why Tool AIs want to be Agent AIs 2016. Available online: https://www.gwern.net/Tool-AI (accessed on 18 February 2019).
63. Yudkowsky, E. Harry Potter and Method of Rationality. 2010. Available online: https://fanlore.org/wiki/Harry_Potter_and_the_Methods_of_Rationality (accessed on 18 February 2019).
64. Bostrom, N.; Armstrong, S.; Shulman, C. Racing to the Precipice: a Model of Artificial Intelligence Development. *AI Soc.* **2013**, *31*, 201–206.
65. Shulman, C. Arms races and intelligence explosions. In *Singularity Hypotheses*; Springer: Berlin, Germany, 2011.
66. Bostrom, N. Strategic Implications of Openness in AI Development. *Glob. Policy* **2016**, *8*, 135–148. [CrossRef]
67. Baum, S.D. On the Promotion of Safe and Socially Beneficial Artificial Intelligence. *Glob. Catastroph. Risk.* **2016**, *32*, 543–551. [CrossRef]
68. Ouagrham-Gormley, S.B. Dissuading Biological Weapons. In *Proliferation Pages*; Springer: Berlin, Germany, 2013; pp. 473–500.
69. Auerbach, D. The Most Terrifying Thought Experiment of All Time. Available online: http://www.slate.com/articles/technology/bitwise/2014/07/roko_s_basilisk_the_most_terrifying_thought_experiment_of_all_time.html (accessed on 18 February 2019).
70. Fernando, C. PathNet: Evolution Channels Gradient Descent in Super Neural Networks. *arXiv*, 2017; arXiv:1701.08734.
71. Nelson, R. How to Deter a Rogue AI by Using Your First-mover Advantage. 2007. Available online: http://www.sl4.org/archive/0708/16600.html (accessed on 18 February 2019).
72. Kharpal, A. *Elon Musk: Humans Must Merge with Machines or Become Irrelevant in AI Age*; CNBC: Englewood Cliffs, NJ, USA, 2017.
73. Brin, D. *The Transparent Society*; Perseus Book: New York, NY, USA, 1998.
74. Hanson, R. *The Age of Em: Work, Love, and Life when Robots Rule the Earth*; Oxford University Press: Oxford, UK, 2016.
75. Bostrom, N. *Hail Mary, Value Porosity, and Utility Diversification*; Oxford University Press: Oxford, UK, 2016.
76. Lem, S. The Investigation. 1959. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1651-2227.1959.tb05423.x (accessed on 18 February 2019).
77. Urban, T. Neuralink and the Brain's Magical Future. 2017. Available online: https://waitbutwhy.com/2017/04/neuralink.html (accessed on 18 February 2019).
78. Bostrom, N. Human genetic enhancements: a transhumanist perspective. *J. Value Inq.* **2003**, *37*, 493–506. [CrossRef]
79. Smith, P.D. *Doomsday Men: The Real Dr. Strangelove and the Dream of the Superweapon*; St. Martin's Press: New York, NY, USA, 2007.
80. Alexander, S. Should AI Be Open. Available online: https://slatestarcodex.com/2015/12/17/should-ai-be-open/ (accessed on 18 February 2019).
81. Baker, B.H. *The Gray Matter: The Forgotten Story of the Telephone*; Telepress: Kent, WA, USA, 2000; ISBN 0-615-11329-X.
82. The Telegraph Russian Spacecraft Landed on Moon Hours Before Americans. Available online: http://www.telegraph.co.uk:80/science/space/5737854/Russian-spacecraft-landed-on-moon-hours-before-Americans.html (accessed on 18 February 2019).
83. Venture Scanner Artificial Intelligence Q1 Update in 15 Visuals 2016. Available online: https://www.venturescanner.com/blog/2016/artificial-intelligence-q1-update-in-15-visuals (accessed on 18 February 2019).
84. Yampolskiy, R. From Seed AI to Technological Singularity via Recursively Self-Improving Software. *arXiv* **2015**, arXiv:1502.06512.
85. Drexler, E.; Phoenix, C. Safe exponential manufacturing. *Nanotechnology* **2004**, *15*, 869.
86. Bontchev, V. *Are Good Computer Viruses Still a Bad Idea?* EICAR: London, UK, 1994.
87. Sotala, K.; Valpola, H. Coalescing minds: brain uploading-related group mind scenarios. *Int. J. Mach. Conscious.* **2012**, *4*, 293–312. [CrossRef]
88. Batin, M.; Turchin, A.; Markov, S.; Zhila, A.; Denkenberger, D. Artificial Intelligence in Life Extension: From Deep Learning to Superintelligence. *Inform. Slov.* **2018**, *41*, 401.
89. Alexander, S. Book Review: Age of Em. Available online: http://slatestarcodex.com/2016/05/28/book-review-age-of-em/ (accessed on 18 February 2019).
90. Bostrom, N. Are You Living in a Computer Simulation? *Publ. Philos. Q.* **2003**, *53*, 243–255. [CrossRef]

91. Omohundro, S. The basic AI drives. In Proceedings of the AGI Frontiers in Artificial Intelligence and Applications, Memphis, TN, USA, 1–3 March 2008.
92. Bostrom, N. Existential risk prevention as global priority. *Glob. Policy* **2013**, *4*, 15–31. [CrossRef]
93. Shakirov, V. Review of State-of-the-Arts in Artificial Intelligence with Application to AI Safety Problem. *arXiv* **2016**, arXiv:1605.04232.
94. DeepMind AlphaGo. Available online: https://deepmind.com/research/alphago/ (accessed on 18 February 2019).
95. Ministry of National Defense of the People's Republic of China. *The Dawn of the Intelligent Military Revolution*; Ministry of National Defense of the People's Republic of China: Beijing, China, 2016.
96. Factored Cognition (May 2018) Ought. Available online: https://ought.org/presentations/factored-cognition-2018-05 (accessed on 25 January 2019).