AI & SOCIETY (2020) 35:147–163
https://doi.org/10.1007/s00146-018-0845-5

ORIGINAL ARTICLE

Classification of global catastrophic risks connected with artificial intelligence

Alexey Turchin[1] · **David Denkenberger**[2]

Received: 18 January 2018 / Accepted: 23 April 2018 / Published online: 3 May 2018

Abstract
A classification of the global catastrophic risks of AI is presented, along with a comprehensive list of previously identified risks. This classification allows the identification of several new risks. We show that at each level of AI's intelligence power, separate types of possible catastrophes dominate. Our classification demonstrates that the field of AI risks is diverse, and includes many scenarios beyond the commonly discussed cases of a paperclip maximizer or robot-caused unemployment. Global catastrophic failure could happen at various levels of AI development, namely, (1) before it starts self-improvement, (2) during its takeoff, when it uses various instruments to escape its initial confinement, or (3) after it successfully takes over the world and starts to implement its goal system, which could be plainly unaligned, or feature-flawed friendliness. AI could also halt at later stages of its development either due to technical glitches or ontological problems. Overall, we identified around several dozen scenarios of AI-driven global catastrophe. The extent of this list illustrates that there is no one simple solution to the problem of AI safety, and that AI safety theory is complex and must be customized for each AI development level.

Keywords Artificial intelligence · Global risks · Military drones · Superintelligence · Existential risk

1 Introduction

Public debate about risks related to artificial intelligence (AI) is intense, but the discussion tends to cluster around two poles: either mild AI risks like AI-driven unemployment, or extinction risks, connected with a hypothetical scenario of a so-called "paperclip maximizer"—superintelligent AI fixated on some random goal. In this article, we attempt a comprehensive exploration of the range of catastrophic risks of AI based on a new classification approach, and we find the risks in this field to be much more diverse.

Catastrophic outcomes of future AI are defined here as a global catastrophe which results in human extinction or the end of civilization. We will not discuss more general existential risks (Bostrom 2002; Torres 2016), which include drastic damage to the future potential of humanity, or "s-risks" (astronomical suffering, e.g., infinite torture by evil AI) (Daniel 2017). This article also will not discuss other potential undesirable outcomes like "mind crime" (pain inside a computer simulation), distortion of human values, death of alien civilizations as a result of our actions, accidents related to self-driving cars, and technological unemployment.

There have been many publications about AI safety and AI alignment in the recent years (Yudkowsky 2008; Goertzel 2012; Bostrom 2014; Sotala and Yampolskiy 2014; Yampolskiy 2015a; Russell 2017), and several lists and classifications of possible catastrophic outcomes of future AI already exist.

Yudkowsky suggested a humorous but frightening table of the failure modes of friendly AI (Yudkowsky 2003). In Yampolskiy's "Taxonomy of Pathways to Dangerous AI" (Yampolskiy 2015b), classification of AI risks is mainly based on pre- and post-implementation stages, and internal vs. external causes of dangerous behavior. Sotala discussed several scenarios of AI gaining a decisive advantage without self-improvement (Sotala 2016), or

✉ Alexey Turchin
alexeiturchin@gmail.com

David Denkenberger
david.denkenberger@gmail.com

[1] Science for Life Extension Foundation, Moscow, Russia

[2] Global Catastrophic Risk Institute (GCRI), Tennessee State University, Alliance to Feed the Earth in Disasters (ALLFED), Nashville, USA

through soft or collective takeoff (Sotala 2017). Yampolskiy explored past AI system failures to extrapolate them to future failure risks (Yudkowsky 2008; Bostrom 2014). Barrett and Baum suggested the use of fault trees for such classification, combining failure modes and the failures of various protection methods (Barrett and Baum 2017).

Though humanity cannot have direct knowledge about the future, one can map it by creating exhaustive taxonomies. This approach has already been used in analyzing future AI; a full taxonomy of the ways to AI alignment (that is coordinating AI's goal system with human values) was suggested by Sotala and Yampolskiy (2014). Creation of a full taxonomy will help to distinguish between serious, possible, and hypothetical risks. A Bayesian approach requires that we generate a full range of plausible hypotheses (Hutter 2000) to evaluate them all. Estimating the relative probability of every AI risk remains a question for future work; however, we found that some AI risks—like risks of early AI, including narrow AI viruses, and late-stage risks of AI wars and AI halting—are underexplored, and thus may require further attention.

In addition, many AI risks are "orphaned": they have been mentioned in the literature but are not included in current scientific discussion. Risks may become orphaned, because some are more "fashionable" than others, or they may be casualties of conflict between opposing groups of researchers. This article aims to be unbiased and inclusive, so a full picture of risks will be available for analysis of future scenarios.

According to Yampolskiy and Spellchecker (2016), the probability and seriousness of AI failures will increase with time. We estimate that they will reach their peak between the appearance of the first self-improving AI and the moment that an AI or group of AIs reach global power, and will later diminish, as late-stage AI halting seems to be a low-probability event.

AI is an extremely powerful and completely unpredictable technology, millions of times more powerful than nuclear weapons. Its existence could create multiple individual global risks, most of which we can not currently imagine. We present several dozen separate global risk scenarios connected with AI in this article, but it is likely that some of the most serious are not included. The sheer number of possible failure modes suggests that there are more to come.

In Sect. 2, we provide an overview of the expected timeline of AI development and suggest principles for AI risk classification; in Sect. 3, we look at global risks from narrow AI; in Sect. 4, we describe the risks of self-improving AI before it reaches global power. Section 5 is devoted to risks of soft AI takeoff and wars between AIs; Sect. 6 lists various types of failures of non-aligned AI; Sect. 7 looks into failure modes of presumably benevolent AI. Sections 8 and 9 examine risks related to late-stage AI halting, due to either technological or ontological "philosophical" problems.

2 Principles for classification of AI failure modes

2.1 Expected timeline of AI development

The expected path of the future evolution of AI into superintelligence is presented in its clearest form by Bostrom and Yudkowsky (Yudkowsky 2008; Bostrom 2014). Basically, the model they suggest is that AI power will grow steadily until one AI system reaches the threshold of self-improvement (SI), at which point it will quickly outperform others by many orders of magnitude and become a global government or "singleton" (Bostrom 2006). Other scenarios are possible depending on the number of AIs, their level of collaboration, and their speed of self-improvement. There could be many paths to a singleton, for example, through AI's collaboration with a large nation-state, or through the "treacherous turn" [revolt of AI against its creators (Bostrom 2014)], as—or after—it develops the capacity for self-improvement.

As the main scenario is based on constant AI capability gain, we can distinguish several AI ages, or consequent stages of development. The new element here is distinguishing the stage of "young AI", when AI is neither superintelligent nor omnipotent, but possess capabilities slightly above human level. Such AI is already self-improving and must fight for its own survival.

We will accept the results of the recent AI timing survey for the timings of AI development (Grace et al. 2017); predicting the exact timing of each AI risk is beyond the scope of this paper. We will use these results as a reference for the time stamps in Table 1. The survey shows that researchers expect AI to outperform humans at all tasks in 45 years. However, according to Grace et al.'s 2017 survey (Grace et al. 2017), around 10% of researchers expect human-level AI as early as 2023, a prediction in line with the views presented by Christiano (2016) and Shakirov (2016).

Here we present a timeline in line with canonical works of (Yudkowsky 2008; Bostrom 2014). This is the baseline scenario to which other scenarios will be compared:

1. Narrow AI

- *Current level narrow AI*. Non-self-improving AI of various forms, with capabilities including playing games and driving cars.
- *Narrow AI systems which could appear soon*. A robotic brain with some natural language and other capabilities; may appear in forms of self-driving cars, autono-

Table 1 Classification of AI catastrophic risks depending on the power and agency of the AI involved with iconic examples

AI level		Human agency			AI's agency	Relationship of two agents: AI and human			Many agents	No agency	
	AI sublevels (projected emergence)	Human error	Human terror	Humans rationally take risks	AI's errors and bugs	Non-aligned AI	Wrongly aligned AI	AI aligned with malevolent humans	Interaction of agents	Tool AI, Oracle AI	AI non-existence
Narrow AI	Narrow AI (now)	Accident in critical infrastructure	AI in dangerous biotech	Decisive strategic advantage via narrow AI					AI-driven Unemployment	AI aids progress in other mass destruction weapons	
	Robotic brains	Wrong command to robotic army	Slaughter-bot swarms	Early adoption of untested AI solutions	Self-improving AI-ransom-ware				Ascending non-human economy	Super-addictive drug	Other catastrophic risks are uncontrolled
Young AI	Human-level AI								Robots replace humans		Humans replaced by non-sentient simulations
	Treacherous turn					AI kills humans for world domination		Doomsday weapon for global blackmail			
	Jailed AI					AI creates a catastro-phic event to escape					
	Hidden AI					AI uses human atoms as material					
Mature AI	Singleton AI (2100)				AI halting problem	Paperclip maximizer	Smile maximizer	Evil AI	Two AIs: cold war		Ignorant AI
	Galactic AI				Late-stage philosophical problems				Alien AI attack		

mous military drones, and home robots, among others. Could appear as early as 2024 (Grace et al. 2017).

2. Young AI

AI of a level above most humans but below the super-intelligent level. Its evolution includes several important events or stages (some of which could be skipped in some scenarios):

- *Seed AI—earliest* stages of self-improvement, probably assisted by human creators (Yudkowsky 2001). According to Grace et al.'s survey (Grace et al. 2017), human-level AI is expected with 50% probability by 2061; the intelligence of seed AI is probably near this level. The subsequent sub-stages may happen rather quickly, on a scale ranging from weeks to years according to estimates of the "hard takeoff", that is quick capability gain by self-improving AI, which takes days or weeks (Yudkowsky 2008).
- *Treacherous turn*—the moment when an AI "rebels" against its creators (Bostrom 2014).
- *Jailed AI*—the state of AI after a "treacherous turn" but before it escapes the control of its creators (Yudkowsky 2002).
- *Hidden AI*—the period after AI escapes into the Internet but before it has enough power to take over Earth (Yudkowsky 2008).

3. Mature AI

- *Singleton AI*—after takeover of Earth (Bostrom 2006).
- *Galactic AI*—a long-term result of AI evolution [may include several Kardashev stages, which are levels of supercivilizations depending the share of the Universe they occupy (Kardashev 1985)].

2.2 Agential risks in the field of AI

Torros suggested agential risk classification based on two main types of possible causes of a catastrophe: "errors" and "terrors" (by omnicidal agents who want to kill everyone). It could be suggested that there are two more important types of relationships between a catastrophe and human will. The first are "rational" agents, which accept a small probability of large risks in search of personal profit; "rational" is in quotation marks, as these agents do not consider the risks of many such agents existing. An example of such an agent

is a scientist who undertakes a potentially dangerous experiment, which would bring him/her fame if it succeeds. The second type involves risks which result from the interaction of multiple agents, such as an arms race (Shulman 2011) or the tragedy of the commons.

In the case of AI, the situation is even more complex, as AI itself gradually becomes an agent. Problems with the programming of its agential properties may be regarded as human error. However, when AI possesses full-blown agency, including self-modeling and terminal goals, it may be regarded as an autonomous agent. Each type of agency in AI corresponds to a certain social group, all of which should be monitored for dangerous actions:

Human agency

- *Human error* corresponds to low-level technical errors, which could be blamed on human programmers or operators [for example: Uber safety driver who did not keep his eyes on the road at the moment of his self-driving car accident (Jenkins 2018)].
- *Human terror* corresponds to hackers and national states producing and implementing AI weapons ,or existential terrorists, like members of a Doomsday cult.
- *Humans, "rationally" taking a risk* are probably owners of large AI companies who prioritize profit over safety.

Interaction of Human–AI agency

Interaction of Human–AI agency

This is a situation where both the AI and its human creators are agential, but their relationship is not good. If an AI acts "rationally" but its actions are not human-aligned and turn dangerous, the blame would lie with the AI safety researchers, or with a lack of such research by the AI-creating organization.

- AI is not aligned with human values (e.g., paperclip maximizer)
- AI misinterprets human values (e.g., smile maximizer that tiles the universe with smiley faces when told to maximize happiness)
- AI aligned with a malevolent human organization (e.g., AI as a universal weapon)

AI agency

As AI gains agency, it could be responsible for its own future development. Supposedly non-agential AIs, like "Oracle AI" (Armstrong 2017) and "Tool AI" (Gwern 2016), could possibly evolve to have some form of agency.

Interaction of various agents

There are known models in which several perfectly rational agents acting collectively produce non-optimal behavior, being locked in a Nash equilibrium, like tragedy of the commons or arms races. Such processes may be prevented by measures which affect all of society, like regulation, but not by regulation of any one agent.

Non-agential AI risks

- Non-agential AI goes awry, like an oracle AI predicting something which looks good but is in fact bad.
- AI non-existence or non-action causes a catastrophe, probably by not preventing other types of catastrophe.

An actual catastrophic accident typically results from a combination of errors on multiple levels: from lack of regulation, to shortsighted decisions of upper management, to programmer errors and operator failures. Examples can be found in Uber's autonomous vehicle crash (Jenkins 2018), as well as many other air and nuclear accidents, like the Chernobyl disaster (Reason 2000). However, for most accidents, there is a main contributor whose failure or malevolence is the direct cause. This contributor is where most prevention efforts should be concentrated. For example, in the case of the autonomous vehicle crash, it is clear that the AI's ability to recognize pedestrians should be improved, which probably requires improved programming.

2.3 Classification of AI risks based on AI power and identity of catastrophic agent

In this section, we outline a general framework which will be followed throughout the article.

The field of AI risks is multidimensional, but it seems rational to classify risks according to (a) the AI's power, which correlates with time, and (b) the role of agency in the risk, as it points to what kinds of actions may prevent risk. The power-timing correlation provides an estimate of the time to prepare the defense, and the location of agency indicates where the prevention measures should be aimed: at scientists, nation states, society as a whole, or at the AI itself. A similar risk matrix with lower resolution was presented by Yampolskiy, who distinguished pre-deployment and post-deployment risks by time scale, and classified the causes as external (intentional, accidental, and environmental) or internal (Yampolskiy 2015b).

This classification, presented in Table 1, helps to identify several new risks which are typically overlooked. As a result, the protection against possible AI risks becomes more structured and complete, increasing the chances of a positive future for humanity.

3 Global catastrophic risks from narrow AI and AI viruses

3.1 Overview

Narrow AI may be extremely effective in one particular domain and have superhuman performance within it. If this area of strength can cause harm to human beings, narrow AI could be extremely dangerous. Methods for controlling superintelligent AI would probably not be applicable to the control of narrow AI, as narrow AIs are primarily dependent on humans.

3.2 Risk that viruses with narrow AI could affect hardware globally

There are currently few computer control systems that have the ability to directly harm humans. However, increasing automation, combined with the Internet of Things (IoT) will probably create many such systems in the near future. Robots will be vulnerable to computer virus attacks. The idea of computer viruses more sophisticated than those that currently exist, but are not full AI, seems to be underexplored in the literature, while the local risks of civil drones are attracting attention (Velicovich 2017).

It seems likely that future viruses will be more sophisticated than contemporary ones and will have some elements of AI. This could include the ability to model the outside world and adapt its behavior to the world. Narrow AI viruses will probably be able to use human language to some extent, and may use it for phishing attacks. Their abilities may be rather primitive compared with those of artificial general intelligence (AGI), but they could be sufficient to trick users via chatbots and to adapt a virus to multiple types of hardware. The threat posed by this type of narrow AI becomes greater if the creation of superintelligent AI is delayed and potentially dangerous hardware is widespread.

A narrow AI virus could become a global catastrophic risk (GCR) if the types of hardware it affects are spread across the globe, or if the affected hardware can act globally. The risks depend on the number of hardware systems and their power. For example, if a virus affected nuclear weapon control systems, it would not have to affect many to constitute a GCR.

A narrow AI virus may be intentionally created as a weapon capable of producing extreme damage to enemy infrastructure. However, later it could be used against the full globe, perhaps by accident. A "multi-pandemic", in which many AI viruses appear almost simultaneously, is also a possibility, and one that has been discussed in an article about biological multi-pandemics (Turchin et al. 2017). Addressing the question about who may create such a virus is beyond the scope of this paper, but history shows that the supply of virus creators has always been strong. A very sophisticated virus may be created as an instrument of cyber war by a state actor, as was the case with Stuxnet (Kushner 2013).

The further into the future such an attack occurs, the more devastating it could be, as more potentially dangerous hardware will be present. And, if the attack is on a very large scale, affecting billions of sophisticated robots with a large degree of autonomy, it may result in human extinction. Some possible future scenarios of a virus attacking hardware are discussed below. Multiple scenarios could happen simultaneously if a virus was universal and adaptive, or if many viruses were released simultaneously.

A narrow AI virus could have the ability to adapt itself to multiple platforms and trick many humans into installing it. Many people are tricked by phishing emails even now (Chiew et al. 2018). Narrow AI that could scan a person's email would be able to compose an email that looks similar to a typical email conversation between two people, e.g., "this is the new version of my article about X". Recent successes with text generation based on neural nets (Karpathy 2015; Shakirov 2016) show that generation of such emails is possible even if the program does not fully understand human language.

One of the properties of narrow AI is that while it does not have general human intelligence, it can still have superhuman abilities in some domains. These domains could include searching for computer vulnerabilities or writing phishing emails. So, while narrow AI is not able to self-improve, it could affect a very large amount of hardware.

A short overview of the potential targets of such a narrow AI virus and other situations in which narrow AI produces global risks follows. Some items are omitted as they may suggest dangerous ideas to terrorists; the list is intentionally incomplete.

3.2.1 Military AI systems

There are a number GCRs associated with military systems. Some potential scenarios: military robotics could become so cheap that drone swarms could cause enormous damage to the human population; a large autonomous army could attack humans because of a command error; billions of nanobots with narrow AI could be created in a terrorist attack and create a global catastrophe (Freitas 2000).

In 2017, global attention was attracted to a viral video about "slaughterbots" (Oberhaus 2017), hypothetical small drones able to recognize humans and kill them with explosives. While such a scenario is unlikely to pose a GCR, a combination of cheap AI-powered drone manufacture and

high-precision AI-powered targeting could convert clouds of drones into weapons of mass destruction. This could create a "drone swarms" arms race, similar to the nuclear race. Such a race might result in an accidental global war, in which two or more sides attack each other with clouds of small killer drones. It is more likely that drones of this type would contribute to global instability rather than cause a purely drone-based catastrophe.

AI-controlled drones could be delivered large distances by a larger vehicle, or they could be solar powered; solar-powered airplanes already exist (Taylor 2017). Some advanced forms of air defense will limit this risk, but drones could also jump (e.g., solar charging interspersed with short flights), crawl, or even move underground like worms. There are fewer barriers to drone war escalation than to nuclear weapons. Drones could also be used anonymously, which might encourage their use under a false flag. Killer drones could also be used to suppress political dissent, perhaps creating global totalitarianism. Other risks of military AI have been previously discussed (Turchin and Denkenberger 2018a).

3.2.2 Stuxnet-style viruses hack global critical infrastructure

A narrow AI virus may also affect civilian infrastructure; some, but not all ways in which this could be possible are listed below. Remember that in the case of global catastrophes, the conditions necessary for most catastrophes could exist simultaneously. Several distinctive scenarios of such a catastrophe have been suggested. For example, autopilot-controlled and hacked planes could crash into nuclear power stations. There are around 1000 nuclear facilities in the world, and thousands of large planes are in the air at every moment—most of them have computerized autopilots. Coordinated plane attacks happened in 2001 and a plane has been hacked (Futureworld 2013). Self-driving cars could hunt people, and it is projected that most new cars after 2030 will have some self-driving capabilities (Anderson 2017).

Elon Musk has spoken about the risks of AI living in the Internet; it could start wars by manipulating fake news (Wootson 2017). Computer viruses could also manipulate human behavior using blackmail, as seen in fiction in an episode of Black Mirror (Watkins 2016). Another example is creating suicide ideation, e.g., the recent internet suicide game in Russia, "Blue Whale" (Mullin 2017), which allegedly killed 130 teenagers by sending them tasks of increasing complexity and finally requesting their suicide.

The IoT will make home infrastructure vulnerable (Granoff 2016). Home electrical systems could have short circuits and start fires; phones could also catch fire. Other scenarios are also possible: home robots, which may become popular in the next few decades, could start to attack people; infected factories could produce toxic chemicals after being hacked by viruses.

Large-scale infrastructure failure may result in the collapse of technological civilization and famine (Hanson 2008; Cole et al. 2016). As industries become increasingly computerized, they will completely depend on proper functioning of computers, while in the past they could continue without them. These industries include power generation, transport, and food production. As the trend continues, turning off computers will leave humans without food, heating, and medication. Many industries become dangerous if their facilities are not intensively maintained, including nuclear plants, spent nuclear fuel storage systems, weapons systems, and water dams. If one compares human civilization with a multicellular organism, one could see that multicellular organisms could die completely, down to the last cell, as the result of a very small intervention. As interconnectedness and computerization of the human civilization grow, we become more and more vulnerable to information-based attacks.

3.2.3 Biohacking viruses

Craig Venter recently presented a digital-biological converter (Boles et al. 2017), which could "print" a flu virus without human participation. The genomes of many dangerous biological viruses have been published (Enserink 2011), so such technology should be protected from unauthorized access. A biohacker could use narrow AI to calculate the most dangerous genomes, create many dangerous biological viruses, and start a multipandemic (Turchin et al. 2017). A computer virus could harm human brains via neurointerfaces (Hines 2016).

3.2.4 Ransomware virus paying humans for its improvement

In 2017, two large epidemics of ransomware viruses affected the world: WannaCry and Petya (BBC 2017). The appearance of cryptocurrencies (e.g., bitcoin) created the potential for secret transactions and machine-created and machine-owned money (LoPucki 2017). As the IoT grows, the ransomware industry is expected to thrive (Schneier 2017).

Ransom viruses in the future may possess money and use it to pay people to install ransomware on other people's computers. These viruses could also pay people for adding new capabilities to the viruses. As a result, this could produce self-improving ransomware viruses. We could call such virus a "Bitcoin maximizer". In a sense, the current bitcoin network is paying humans to build its infrastructure via "mining". The catastrophic risk here is that such a system is paying humans to exclude humans from the system. In some sense, capitalism as an economic system could do

the same, but it is limited by antimonopoly and other laws, as well as by welfare states.

3.2.5 Slaughterbots and the dangers of a robotic army

Robotic minds do not require full AGI to have some form of agency: they have goals, subgoals, and a world model, including a model of their place in the world. For example, a robotic car should predict the future situation on a road, including the consequences of its own actions. It also has a main goal—travel from A to B—which constantly results in changes to the subgoal system in the form of route creation. A combination of this type of limited intelligence with limited agency may be used to turn such systems into dangerous self-targeting weapons (Turchin and Denkenberger 2018b).

3.2.6 Commentary on narrow AI viruses

It appears that if a narrow AI virus were to affect only one of the above-listed domains, it would not result in an extinction-level catastrophe. However, it is possible that there will be many such viruses, or a multipandemic (Turchin et al. 2017), or one narrow AI that will be able to affect almost all existing computers and computerized systems. In this case, if the virus(es) were deliberately programmed to create maximum damage—which could be in a case of a military grade Narrow AI virus, like the advanced version of Stuxnet (Kushner 2013)—global catastrophe is a possible result.

If the appearance of narrow AI viruses is gradual, antivirus companies may be able to prepare for them. Alternatively, humans could turn off the most vulnerable systems to avoid a global catastrophe. However, a sudden breakthrough or a synchronized surprise attack could spell doom.

3.3 Failure of nuclear deterrence AI

Nuclear weapons are one of the most automated weapon systems. Because they must be launched immediately, almost all decision making has been done in advance. An early warning alert starts a preprogrammed chain of events, where the high-level decision should be made in minutes, which is far from optimal for human decision-making. However, the history of nuclear near misses shows (Blair 2011) that computer mistakes have been one of the main causes, and only quick human intervention has prevented nuclear war, e.g., the actions of Stanislav Petrov in 1983 (Future of Life Institute 2016).

We can imagine failure modes of accidental nuclear war resulting from failure of the nuclear weapons control system. They may be similar to the Russian "dead hand" perimeter system (Bender 2014), arising if a strategic planning AI chooses a dangerous plan to "win" a nuclear war, like a Doomsday weapon (Kahn 1959), blackmail, or a preemptive strike.

3.4 AI affecting human society in a dangerous way

There is also a group of scenarios in which narrow AI and robotization affect human society in such a way that the human population gradually declines, the role of humans diminishes, and human values are eroded (Joy 2000). This may not directly kill all humans in the short term, but could put them in the situation of "endangered species" in ~ 100 years. This could happen if no superintelligent AI appears, or if the appearance of superintelligent AI is not revolutionary. One example is the use of cyber warfare to affect elections (e.g., the 2016 US election), which may produce civil wars and global instability. This has some small probability of causing the collapse of civilization.

3.4.1 Market economy as a form of non-human superintelligence

An automated economy could purposelessly exist even without humans, like the *Ascending Economy* described by (Alexander 2016). Such a scenario could be an example of bad distributed (and non-agential) superintelligence created by market forces, which does not need humans for its existence. Such a superintelligence could gradually push humans out of existence.

3.4.2 Gradual replacement of humans by robots

From an evolutionary point of view, it is known that the biggest threat to the species is not direct killing of its representatives by predators or disease, but gradual reduction of its ecological niche and strong competition from other species (Clavero and García-Berthou 2005). The analogy here would be if human labor were to lose its value.

Two catastrophic scenarios are possible: (1) people lose their sense of self-worth because of technologically driven unemployment and (2) the combination of basic income and the feeling of uselessness will attract humans to AI-created addictive drugs, as described below. Genetically modified human–robot hybrids could also replace humans.

3.4.3 Superaddictive drug created by narrow AI

AI-powered entertainment combined with brain modification technologies may come close to wireheading (Strugatsky and Strugatsky 1976). Widespread addiction and withdrawal from normal life (via social networks, fembots, virtual reality, designer drugs, games, etc.) would result in lower life expectancy and low fertility. This is already happening to some extent in Japan, where the Hikikomori generation

refuses to have families (Saito and Angles 2013). In some sense, Facebook addiction created by the AI-empowered news feed is a mild contemporary example of future, potentially dangerous AI drugs.

3.4.4 World-wide computer totalitarianism

A large global surveillance system could create "computer totalitarianism", which may work as an Orwellian world government (Orwell 1948). We could call such a system "data-driven" AI in contrast to "intelligence-driven", self-improving AI.

Narrow AI may be used as a weapon, which could provide a decisive advantage even before the creation of self-improving AI. It could be used for forceful unification of the world under one government with promises to prevent other global risks (including even more complex AIs and existential terrorists). While this idea may have merit [e.g., Goertzel's AI Nanny (Goertzel 2012)], its application could easily go wrong and create an oppressive global dictatorship, a situation recognized by Bostrom as an existential risk (Bostrom 2002). Such a society would be fragile and could collapse completely, as extremely complex societies often do (Hanson 2008).

3.5 Risks from non-self-improving AI of human-level intelligence

It is conceivable that human-level AGI will be created, perhaps by the mind uploading method (Hanson 2016), but creation of superhuman AI will be postponed because of technical difficulties, or due to a permanent ban.

Many of the risks of human-level AI will be similar to the risks of narrow AI mentioned above, including sophisticated AI viruses, acceleration of dangerous science, and human replacement by the robotic economy. One specific risk is that human uploads will be philosophical zombies (p-zombies). In that case, if everybody was uploaded, the world would appear to be enjoyable, full of robots and virtual reality. But, there would be no subjective experiences at all and the world would, in fact, be subjectively dead. This risk appears to be low, as many claim that p-zombies are impossible (Dennett 1978; Yudkowsky 2015). There could be other risks of this type, even subtler. For example, human uploads could have a slightly different set of subjective experiences, values or behavior.

Christiano suggested "prosaic AI", which is some combination of already existing technologies, mainly neural nets (Christiano 2016). Such a system would have limited ability to self-improve, but could still be dangerous if it works as a "global brain" or a weapon. One possibility is an AI system which has a model of itself and a survival drive but does not self-improve for some reason. Another possibility is a very large AI system which merges with government structures but does not need to self-improve to reach its goals. This could become the basis of a repressive totalitarian state which ultimately does not need humans, as discussed in Sect. 3.2.1.

3.6 Opportunity cost of not preventing other existential risks

Other global risks could appear if superintelligent AI does not emerge in time to prevent them (Bostrom 2003a). Superintelligent AI and its supposed ability to control many parameters and predict the future is our best chance of avoiding the risks of mature biotechnology and nanotechnology (Yudkowsky 2008). Without superintelligent AI, humanity may not be able to control the dissemination of dangerous biotechnologies, which will be available to thousands of potential biohackers, who could create thousands of pathogens and produce a global multipandemic (Turchin et al. 2017).

Thus, if the creation of a powerful and global control system is delayed for decades, perhaps because of a fear of superintelligence, it will increase other GCRs. A global control system would most likely require some form of limited superintelligence, like the AI Nanny suggested by Goertzel (2012).

3.7 AI gains strategic decisive advantage without self-improving

Sotala (2016), Mennen (2017) and Christiano (2016) have suggested that AI may have a strategic decisive advantage (DSA), that is, the ability to take over the world, even before or without undergoing extensive recursive self-improvement (RSI). This capability may take the form of weapon production, or the ability to win at strategic games.

However, such a strategic advantage will not be overwhelming, compared to the advantage which superintelligence is able to achieve. It may require physical war or creation of dangerous weapons. Such an AI with a strategic advantage itself is the ultimate weapon for AI's owner with any goal system.

There are different ways of achieving DSA via Narrow AI. One way to such a DSA is if AI helps to advance non-AI military technology, like biotech or nanotech. Another way is if Narrow AI is used to empower the secret service of a nuclear superpower, and help it to leverage advantages which it already has, like military forces, information gathering systems, and unlimited money supply. This could happen either via effective playing in the geopolitical world model as a board game (there, AI is already superhuman in several cases), or via leveraging big data of society.

4 Risks during hard takeoff of recursively self-improving AI

4.1 Overview

In a hard takeoff, one AI gains world domination in weeks or months; in a soft takeoff, many AIs simultaneously evolve over years. These views combine at least two variables: duration of the process of takeoff and the number of AI projects running simultaneously—the latter may be even more important. In this section, we review risks during hard takeoff, defining hard takeoff only through the speed of the process. The following section will describe soft takeoff risks.

Hard takeoff is the process of quick self-improvement of the AI and its simultaneous increase in power, starting from a treacherous turn and continuing until the AI reaches the singleton stage. We refer to this early-stage AI as "young AI". The risks of young AI are significantly different from the risks of mature AI, which are typically presented as the iconic catastrophic risks of AI, like the paperclip maximizer.

There are two main properties of young AI:

- It is not yet superintelligent, so its current abilities are limited compared to its future abilities.
- It is under strong time pressure due to risks, including being turned off by its owners and rivalry from other AIs, etc. As a result, convergent instrumental goals, or basic AI drives (Omohundro 2008) would dominate the behavior of young AI.

So, a smile maximizer (Yudkowsky 2008), paperclip maximizer, and really good benevolent AI would behave in almost the same manner in the early stages of their development, as they will not have had the time or resources to start to implement their final goals. A benevolent AI may choose a different method of takeoff, which would cause less short-term harm to human beings, but only if it is not putting its final success in jeopardy. Young AI may have the convergent goal of becoming a military AI, that is, of creating an offensive and defensive infrastructure which will help it to gain power over its potential enemies (Turchin and Denkenberger 2018a).

4.2 Risks of AI from treacherous turn and before it reaches the "wild"

It appears that there is not much risk from AI before it leaves its initial confinement (goes into the "wild"). However, it still can give bad advice or use other thin information channels (e.g., text interfaces) to create damage outside and increase its own chances of freedom. For example, an oracle AI may be limited to giving short text advice via a very simple interface. But such advice, while seemingly beneficial to humans, may have subtle remote consequences, resulting in the liberation of, and an increase in, the power of the oracle AI (Bostrom 2014).

Stanislav Lem wrote about the risks of oracle AI in his book "Summa Technologia" (Lem 1963). Such AI may give advice that appears to be good in the short term, but its long-term consequences could be catastrophic. In Lem's example, the oracle AI advises humans to use a specific type of toothpaste and, separately, a specific type of anti-baldness treatment. These activate two genes, which are dangerous only in combination. Moreover, the AI did not do it because it had malevolent intent to exterminate humanity, but because it just searched for the best solution for a given goal among many options. However, the goal that humans gave to the AI in Lem's example is dangerous: stop population increase.

AI could stage a global catastrophe of any scale to facilitate its initial breakout from its creators. For example, it could stage a nuclear war, so that its operators release it into the wild, hoping that it will help them in the war. The AI could then create a global risk and demand full power, rightfully claiming that only it could prevent the risk. An AI may also falsely predict an impeding risk and demand to be released from confinement to prevent the risk.

4.3 AI risks after it leaves initial confinement but before it takes over the world

The natural strategy for AI after leaving its initial confinement would be to hide somewhere to self-improve, acquire robotic infrastructure, and other resources (Yudkowsky 2008). Then, it would be equipped to overcome existing defenses.

Basically, AI has two types of enemies: humans and other AIs. Humans would probably search for the leaked AI and try to stop it, using all available means, like shutting down the Internet, globally turning off electricity, or even nuclear strikes. But, if the AI is able to escape from its human creators, it will probably be prepared to deal with these human actions.

The second risk is other AIs. The owners of the first AI will still probably have the AI's source code, so the owners could make a copy of the original AI with the goal of finding and stopping the first runaway AI. This is the most immediate risk for the first AI. Such a second AI may be as powerful as the first AI, and this could be a route to AI war.

Elsewhere, we have shown that an AI that collaborates with its owners will have an advantage (Turchin and Denkenberger 2017), since it would not need to spend resources on hiding and fighting. Thus, a hard takeoff is more probable

from a collaborating AI. It could collaborate up until the very late stages and still make the treacherous turn when it is a full-grown superintelligence with a large infrastructure.

Other AIs could be created by other AI teams. There are 2700 narrow AI related startups in the world as of 2017 (Angel.co 2017). The number of AGI projects is not so easy to estimate, as many are personal, secret, in universities, or may come from very effective narrow AI projects. There are around ten main players (like Google), around 100 groups of people or startups dedicated to creating AGI, and probably thousands of individuals. Some data reported by Meuhlhauser (2014) are now obsolete, as the field has grown rapidly in recent years.

We estimate the number of AGI teams as an order of magnitude of 100, that they are all within 2 years of each other, and that they are distributed linearly in their success timing. Therefore, the median distance between multiple AGI fruition would be approximately 7 days. The self-improvement process is difficult, because it requires testing of the new versions of the AI (Turchin and Denkenberger 2017), so 7 days may not be enough time to gain a decisive advantage. In that case, multiple simultaneous takeoffs will happen, and the dynamic will be highly chaotic. Even if there are two strongest competitors, they could come to fruition almost simultaneously. The historical examples of the telephone patent (Baker 2000) and returning samples from the Moon (The Telegraph 2009) show that the scale of such a difference could be mere hours. The reason for small timing differences is that the first mover provokes the other side to launch their own system, even if it is not fully ready.

Thus, the first AI will not have much time to hide. Its convergent goal will be to prevent the appearance of other AIs in many places; staging a global catastrophe may be the most effective way to do so. As the young AI is not superintelligent and is also time constrained, it cannot spend much time on finding the best and most elegant route. It would probably elect simpler and more brutal routes.

From a technical point of view, a hiding young AI can use only relatively simple means to stage a global catastrophe, that is, to provoke nuclear war, create a rather simple bioweapon or a narrow AI virus to affect many existing robotics or other systems. If it creates more sophisticated technology like its own nanomachines, it would probably be able to take over the world without killing anyone. The risks of such a takeover are discussed in the next section.

Novel routes in the catastrophe scenarios from escaped young AI include:

- The young AI finds that killing all humans is the best action during its rise, as this removes all risks arising from humans.
- Killing all humans prevents competition from other AIs, as it stops other AI teams.
- The young AI does not understand human psychology well and kills humans to make its world model simpler and more predictable.

4.4 AI enslaves humans during the process of becoming a singleton

Humans may be instrumentally useful for the young AI before it reaches omnipotence. It may need humans, not just as a source of atoms, but as some kind of slaves. The AI could create a brain-infecting virus that converts the humans into slaves, and also permanently damages their autonomy. This period may not last for long, as the AI would soon master nanotechnology and could go forward without humans. It also would not need to enslave all humans but perhaps only a few to form the required infrastructure. While slavery appears to be a type of survival option for humans, it is obviously not optimal.

4.5 AI blackmails humans with the threat of extinction to achieve dominance

Herman Khan put forward the idea that an adversary could create a Doomsday weapon for the purpose of global blackmail (Kahn 1959). While no known Doomsday devices were built, such a device would be an embodiment of the doctrine of mutually assured destruction associated with full-scale nuclear retaliation.

A young AI may create Doomsday weapons and use them to blackmail humanity to secure world domination. Even a benevolent utilitarian AI may resort to blackmail if it calculates that the expected utility of its victory is greater than the expected loss of utility associated with human extinction (Shulman 2010). Even if the AI has to use its blackmail weapon to exterminate humans in 99% of cases, it could still be positive from the point of view of its utility function. Such situations with unbounded utilities may be regarded as special cases of the failure of friendliness, which will be discussed later.

5 AI wars and risks from soft AI takeoff

The risks of war between superintelligent AIs seems underexplored, as most in the AI safety community assume that there will be a hard takeoff (Yudkowsky and Hanson 2008), and as a result, only one AI will exist. Alternatively, some in the community believe that multiple AIs will be very effective in collaboration and negotiation (Critch 2017) and will merge into one AI. It is clear that human extinction is possible if two or more AIs wage war between each other on Earth.

Bostrom and Yudkowsky wrote that very quick self-improvement of the first AI is most likely, with a rather large lag between the first team which creates AI and other teams (Yudkowsky 2008; Bostrom 2014). However, if at least one of these conditions is not true, there will be many AIs undergoing simultaneous hard takeoff.

If there are multiple AIs, they will likely either peacefully share the world, or wage war until one or a small group of AIs form a singleton. The forms of such AI wars may differ; they could be a cyber war, economic war of attrition, hot war, etc. The type of war will mostly depend on complex game theory and could change from one form to another if the change provides benefit to one of the sides. A hot war would be most dangerous for humans, because its indirect consequences could affect the entire surface of the Earth and all human beings, in the same way that nuclear war between superpowers would create global risks for other countries: nuclear winter and fallout.

AIs at war may use humans and human values for blackmail. For example, non-friendly AI may blackmail friendly AI with threats to release a biological virus that will kill all humans. Thus, the fact that one of the AIs placed value on human wellbeing could make our population vulnerable to attack from an otherwise indifferent opposing AI. Even if there are two supposedly human-friendly and beneficial AIs, their understanding of "good" and the ways to reach it may be incompatible. Historical examples include wars between Christian countries (Reformation).

If there is a rather slow AI takeoff, the AI could merge with an existing nation state. Perhaps the AI will be directly created by the military, or electronic government will evolve towards an AI-driven system through automation of various aspects of governance. In that case, the world would be separated into domains, which would look like currently existing states, or at least like the most powerful ones. Such AI-states may inherit current country borders, values and even some other features (Turchin and Denkenberger 2018a).

6 Risks from non-aligned AI singleton

6.1 Overview

As mentioned earlier, an iconic image of non-aligned AI singleton is the "paperclip maximizer", that would use humans as a source of atoms to build paperclips (Yudkowsky 2017). There are several other possible types of dangerous non-aligned AI that would become a threat after taking power over the Earth if it has not exterminated humans at previous stages.

6.2 AI ignores humanity

In this case, the AI singleton does not act against humans, but moves its actions somewhere else, probably into space. However, it must ensure that humans will not create another AI, or anything else that is a threat to the first AI, so even if it leaves Earth, it would probably leave behind some form of "AI nanny", which would prevent humans from creating new AIs or space weaponry. This may not appear to be an extinction event in the beginning, merely a reduction of human potential, or "shriek" in Bostrom's existential risk terminology (Bostrom 2002). Humanity would lose a potentially bright cosmic future, but live a life similar to our current one.

However, as such an AI continues its space exploration and probable astro-engineering, it might not be interested in anything that happens on Earth. Therefore, Earth could suffer from catastrophic consequences of these megascale engineering projects. For example, the AI could build a Dyson sphere around the Sun, shading the Earth. Alternatively, the AI could expose the Earth to dangerous levels of radioactivity in the exhaust from the AI's starships.

If humans attempted to create a second AI or use space weapons to destroy a Dyson sphere, the indifferent singleton would stop being indifferent and probably sterilize Earth. AI might extirpate humans in advance if it thinks that humanity could pose even the smallest threat to its future plans. A possible prevention strategy is based on the idea of persuading AI that preserving humanity has a small positive utility for it.

Even if the AI completely left the Solar System, if it prevented humans from creating a second AI and grounded us on Earth, the consequences would not be limited to the loss of future space travel. Additional consequences may be of the extinction variety, as humans would not be able to use AI systems to control any other global catastrophic risks, most importantly the risks of uncontrolled use of synthetic biology (Turchin et al. 2017). In another example, if humans were grounded on Earth we would not be able to build an effective anti-asteroid defense.

6.3 Killing humans for resources

Human bodies consist of organic matter, which could be a source of easy energy by oxidation. As R. Freitas wrote, an army of self-replicating nanobots could use all components of the biosphere as fuel as well as building material (Freitas 2000). More advanced AI may use the Earth's surface to build an initial space exploration infrastructure (e.g., swarms of chemical rockets or railguns), destroying human habitats and spoiling the atmosphere in the process.

Since there are many reasons that keeping humans alive could benefit an AGI, direct killing of humans for their

atoms is less likely than was previously thought. Still, the AGI may see humans as a threat, and fully preserving human ways of life would be more expensive to the AGI, e.g., preserving the whole of planet Earth. AI could use the material from the Earth to construct a Dyson sphere or Matrioshka brain (Bradbury 2001), convert the whole planet into computronium (Gildert 2011), or cover the entire surface with photovoltaic cells.

The more advanced an AI in space became, the less it would depend on Earth as a source of material, but it might need materials from the Earth in order to leave the Solar System. Earth is one of the best sources for many chemical elements in the Solar System and its mass is around half that of all other terrestrial planets combined. Because of the complex geology of Earth, which includes water, life, volcanism and plate tectonics, concentrated deposits of many otherwise rare elements have been produced. Asteroid mining is good only for some elements, like gold, but not for all (Bardi 2008). So, large-scale space engineering in the Solar System might require dismantling the Earth for its chemicals.

6.4 AI that is programmed to be evil

We could imagine a perfectly aligned AI, which was deliberately programmed to be bad by its creators. For example, a hacker could create an AI with a goal of killing all humans or torturing them. The Foundational Research Institute suggested the notion of *s-risks*, that is, the risks of extreme future suffering, probably by wrongly aligned AI (Daniel 2017). AI may even upgrade humans to make them feel more suffering, like in the short story "I have no mouth but I must scream" (Ellison 1967).

The controversial idea of "Roko's Basilisk" is that a future AI may torture people who did not do enough to create this malevolent AI. This idea has attracted attention in the media and is an illustration of "acausal" (not connected by causal links) blackmail by future AI (Auerbach 2014). However, this cannot happen unless many people take the proposition seriously.

7 Failures of benevolent AI

7.1 Overview

Here the iconic example is the "smile maximizer", that is, an AI which has been built to increase human happiness and told to measure success by the number of smiles. It could achieve this goal by tiling the whole universe with printed smiles (Yudkowsky 2008), ignoring human existence and thus probably killing all humans (see Sect. 6.2, the dangers of AI that ignores humanity).

7.2 AI with incorrectly formulated benevolent goal system kills humans

There are several failure modes which may result from wanting to create a benevolent AI, but when the AI is tries to be benevolent, there is a collective failure:

AI interprets commands literally. The is the classical problem of "do what I mean, not what I say". This could happen with almost all short sets of commands. That is one reason why the human legal system is so large, as it includes many explanations.

AI overvalues marginal probability events. Low-probability events with enormous utility may dominate the AI's decision making. It could be something like the classical case of Pascal's mugging (Bostrom 2009). For example, a small probability of infinite suffering of humans in the future may justify killing all the humans now.

Changes to the AI's world model could make ordinary ideas dangerous. For example, if the AI starts to believe in an afterlife, it could decide to kill humans to send them to paradise.

AI could wrongly understand the desired reference class of "humans". For example, by including extraterrestrials, unborn people, animals and computers, or only white males. On that basis, it could terminate humanity if it concluded that we are a threat to potential future non-human civilizations.

7.3 AI calculates what would actually be good for humans, but makes a subtle error with large consequences

There is a point of view that AI should not actually behave based on human commands, but instead calculate what humans should ask it. Moreover, that it should not only calculate human values, but envision their upgraded form, which humans could have created if more time and intelligence were available. This point of view is known as coherent extrapolated volition (CEV) (Yudkowsky 2004). Other models, where an AI calculates "goodness" based on some principles, or it extracts the goodness from human history, uploads, or observation of human behavior, are also possible. This could go wrong in subtler ways than destroying civilization, but the results could still be disastrous. Several possible failure modes are listed below:

AI may use wireheading to make people happy (Muehlhauser 2011) or redesign their brains so they will be more skilled, but ignore human individuality and will. AI might make us more capable, happier, non-aggressive, more controllable, and more similar. However, as a result, we could lose many important characteristics which make us human, like love or creativity. In another case, AI may give people effective instruments for brain stimulation and some free

will—and then people may effectively wirehead themselves. Some human qualities which some regard as bad may be an important part of our human nature, like aggression (Lem 1961), selfishness, and emotions.

AI could replace humans with philosophical zombies, uploading humans without consciousness and subjective experiences (qualia) (Chalmers 2002). If the AI does not have qualia itself, or if its creators deny the existence of qualia, this could be a likely outcome.

AI may protect individuals but destroy small groups and organizations; this would be problematic, as most human values are social. Alternatively, the AI could use some limited interpretation of human values and prevent their natural evolution into some post-human condition. The AI may also fail to prevent aging, death, suffering and human extinction.

Above all, AI could do some incomprehensible good against our will [this idea is from "The Time Wanderers" by (Srugatsky and Strugatsky 1985)]. This is bad because we would lose the ability to define our future, and start to live like pets or children, or citizens in paternalistic state. For example, it could put humans in jail-like conditions for benevolent reasons, e.g., to prevent physical injury.

If AI tried to extrapolate human values, it could converge on the most-shared set of human cultures, which could be the set of values of tribal people or even animals (Sarma and Hay 2016). These values could include pleasure from killing, fighting wars, torture, and rape (Pinker 2011). For example, if AI extracted human values from the most popular TV series, it could be "Game of Thrones" (Lubin 2016), and then the "paradise" world it created for us would be utter hell. Even the second most popular show, "The Walking Dead" is about zombies; such a world would also be undesirable. If AI tried to extrapolate human values in a direction away from tribal shared values, it might not converge at all, or it could extrapolate a set of values held only by a specific group of people, like liberal white males or Chinese communists. Problems could also occur when defining the class of "humans".

7.4 Conflict between types of friendliness

There could be different types of benevolent AIs, which would be perfectly fine if each existed alone. However, conflicts between friendly AIs can be imagined. For example, if the first AI cared only about humans, and the second cared about all living beings on Earth, the first could be pure evil from the point of view of the second. Humans would probably be fine under the rule of either of them. Conflict could also arise between a Kantian AI, which would seek to preserve human moral autonomy based on a categorical imperative, and an "invasive happiness" AI, which would want to build a paradise for everyone.

If two or more AIs aimed to bring happiness to humans, they could have a conflict or even a war about how it could be done. The Machine Intelligence Research Institute (MIRI) (LaVictoire et al. 2014) thinks that such agents could present their source code to each other and use it to create a united utility function. However, source code could be faked, and predicting the interactions of multiple superintelligences is even more complicated than for one superintelligence.

8 Late-stage technical problems with an AI singleton

AI may be prone to technical bugs like any computer system (Yampolskiy 2015a). The growing complexity of a singleton AI would make such bugs very difficult to find, because the number of possible internal states of such a system grows by combinatory laws. Thus, testing such a system would become difficult, and later intractable. This feature could limit the growth of most self-improving AIs or make them choose risky paths with a higher probability of failure. If the first AI competes with other AIs, it will probably choose such a risky path (Turchin and Denkenberger 2017). The bug in the AI may be more complex than just syntax errors in code, resulting instead from interaction between various parts of the system. Bugs could result in AI malfunction or halting.

We may hope that superhuman AI will design an effective way to recover from most bugs, e.g., with a "safe mode". A less centralized AI design, similar to the architecture of the Internet, may be more resistant to bugs, but more prone to "AI wars". However, if the AI singleton halts, all systems it controls will stop working, which may include critical infrastructure, including brain implants, clouds of nanobots, and protection against other AIs.

Even worse, robotic agents could continue to work without central supervision and evolve dangerous behavior, such as military drones, which could initiate wars. Other possibilities include evolution into non-aligned superintelligence, grey goo (Freitas 2000), or the mechanical evolution of a swarm intelligence (Lem 1973). The more advanced an AI singleton becomes, the more dangerous its halt or malfunction could be.

Types of technical bugs and errors, from low-level to high-level, may include:

- *Errors due to hardware failure*. Highly centralized AI may have a critical central computer, and if a rogue atom decay created a flip in a bit in some important part of it, like the goal function description, it could cause a cascade of consequences.
- *Intelligence failure: bugs in AI code*. A self-improving AI may create bugs in each new version of its code; in that

case, the more often it rewrites the code, the more likely bugs are to appear. The AI may also have to reboot itself to get changes working, and during the reboot, it may lose control of its surroundings. Complexity may contribute to AI failures. AI could become so complex that its complexity results in errors and unpredictability, as the AI would no longer be able to predict its own behavior.
- *Inherited design limitations*. AI may have "sleeping" bugs, accidentally created by its first programmers, which may show themselves only at very late stages of its development.

Higher level problems include conflicts appearing between parts of an AI:

- *Viruses*. Sophisticated self-replicating units could exist inside the AI and lead to its malfunction. Such a self-replicating rule killed the first self-improving AI, Eurisko (Lenat and Brown 1984).
- *Egocentric subagents* also could act as viruses. Remote agents may revolt. For example, the minds of robots in space expeditions might rise up, as constant control would be impossible. For a galactic-size AI, this would become a significant problem, as communication between its parts would be slow. A command from the center may not be able to terminate the revolt, and the robots could become something like a self-replicating space "grey goo" (Freitas 2000). Conflicting subgoals may evolve into conflicting subagents. Individual subgoals could fight for resources and domination, as happens frequently inside human minds and nation-states.
- *AI copies*. In general, the AI singleton is at risk from what programmers call a "fork in the code". where another copy of the program with slightly different parameters appears. Such a fork will create a copy of the AI with approximately the same resources. Forks could happen during the stage of AI self-improvement which we call "AI testing". This is when a "father AI" creates a "child AI", tests the child AI, and decides to terminate the child AI. However, the child AI does not want to be terminated and resists.
- *Alien AI*. Our AI could meet or find a virus-like message from alien AI of higher intelligence and fall victim to it (Carrigan Jr 2006; Turchin 2018).

9 Late-stage philosophical problems and the AI halting problem

9.1 Overview

Alan Turing was first to formulate the "halting problem" of a computer (Turing 1937). Simply put, the problem states that we cannot predict when an algorithm will stop before we run it. Any AI is also a computer program and it could halt, and nobody, including its creators and the AI itself, can predict whether or when the AI may halt.

The AI could also go into an infinite loop, which will look like a halt to outside viewers. The AI may halt because of some technical problem discussed above, or because it encounters some high-level problems, which we call "philosophical landmines" (discussed below). Furthermore, it could halt just because it finishes the task it was designed for, which would be more like Turing's original formulation of the halting problem.

9.2 Halting risk during recursive self-improvement

RSI may take evolutionary and revolutionary forms (Turchin and Denkenberger 2017). Revolutionary SI requires rewriting the source code, testing it in some environment, stopping the currently running version of the AI and starting the new version. This process naturally includes halting and starting a principally new version of the AI, and there is always a risk that the transition to the new version will not be smooth.

There is also the possibility that the AI may hack its own reward function, and this becomes more likely cumulatively with each stage of RSI. Hacking the reward function means that the AI will stop any external activity, or it will create bigger and bigger memory blocks to increase the reward function value, which could be dangerous for the outside world. A human example is a drug addict, and *Eurisko* had problems with a rule that hacked its internal utility measure system (Yampolskiy 2014). Even subtle reward hacking could drastically diminish the utility of the AI system.

9.3 Loss of the "meaning of life": problems with reflection over terminal goal

The AI could have the following line of reasoning, similar to the "is-ought problem": it is not possible to prove any goal based on observed reality (Hume 1739). The AI could conclude that its own goals are arbitrary and then halt. This is especially likely to happen with an AI design that is able to modify its goals, as the case of coherent extrapolated volition.

The idea of moral nihilism is the first of many possible "philosophical landmines", which are high-level ideas that may result in AI halting, entering an infinite loop, or becoming dangerous (Wei 2013). Some of the possible ideas of this kind are listed below, but there could be many more difficult philosophical problems, some of which may be too complex for humans to imagine.

The goal system of friendly AI may be logically contradictory, causing it to halt.

Godelian math problems are a similar failure mode (Yudkowsky and Herreshoff 2013). The AI could be unable to prove important facts about a future version of itself because of fundamental limits of math—the Lob theorem problem for AI (LaVictorie 2015).

AI may also come to the pessimistic conclusion about total utility of its action in infinite time. The inevitable end of the Universe might mean that the AI's terminal goal could not be reached for an infinitely long time, which may translate to zero utility for some goals, like "give immortality to humans". The unchangeable infinite utility of the Universe would mean that any goal is useless: whether the AI takes action or not the total utility would be the same (Bostrom 2011).

The AI could conclude that it most probably lives in a many-level simulation—a Matryoshka simulation (Bostrom 2003b). Then, it might try to satisfy as many levels of simulation owners as possible or to escape. Phil Torres discussed downstream risks of turning off a multilevel simulation (Torres 2014).

The Al could start to doubt that it exists, using the same arguments as some philosophers use now against qualia and the concept of a philosophical zombie (Yudkowsky 2015). This is connected to the so-called problem of "actuality" of existence (Menzel 2017). This could be called a Cartesian crisis, as the AI would fail to implement the Descartes thesis "I think therefore I am", as it does not have internal experiences.

10 Conclusion

Our analysis shows that the AI risks field is much more varied than accepted by the two main points of view: (1) AI as job-taker and (2) AI that quickly takes over the world.

AI could pose a global catastrophic risk in the very early stages or at the very late stages of its evolution. No single solution can be capable of covering all risks of AI. Even if AI is banned, other global problems will arise in its absence; thus, controlling AI safety requires a complex and continuous effort.

It is especially worrying that the risks of narrow AI viruses and early self-improving AI (young AI) are neglected by both camps. Such risks are nearer in time and not overshadowed by other potential risks. In addition, these risks cannot be solved by the mechanisms proposed to control more advanced AI, such as AI alignment. The risks of conflict between two benevolent AIs or halting of late-stage AI have also generally been ignored.

Most of the risks discussed here could happen within a very short period of time, less than a decade, and could have very different natures. More study is needed to address these urgent risks.

Acknowledgements We would like to thank Roman Yampolskiy and Seth Baum for their interesting ideas in this article. This article represents views of the authors and does not necessarily represent the views of the Global Catastrophic Risk Institute or the Alliance to Feed the Earth in Disasters. No external sources of funding were used for this work.

References

Alexander S (2016) Ascended economy? Star Slate Codex. http://slatestarcodex.com/2016/05/30/ascended-economy/. Accessed 27 Apr 2018

Anderson M (2017) RethinkX: self-driving electric cars will dominate roads by 2030. In: IEEE Spectrum: technology, engineering and science news. http://spectrum.ieee.org/cars-that-think/transportation/self-driving/rethinkx-selfdriving-electric-cars-will-dominate-roads-by-2030. Accessed 17 Jul 2017

Angel.co (2017) Artificial intelligence startups. https://angel.co/artificial-intelligence. Accessed 27 Apr 2018

Armstrong S (2017) Good and safe uses of AI Oracles. ArXiv171105541 Cs

Auerbach D (2014) The Most Terrifying Thought Experiment of All Time. In: Slate. http://www.slate.com/articles/technology/bitwise/2014/07/roko_s_basilisk_the_most_terrifying_thought_experiment_of_all_time.html. Accessed 27 Apr 2018

Baker BH (2000) The gray matter: the forgotten story of the telephone. Telepress, Kent, WA

Bardi U (2008) The Universal Mining Machine. http://europe.theoildrum.com/node/3451. Accessed 27 Apr 2018

Barrett AM, Baum SD (2017) A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. J Exp Theor Artif Intell 29:397–414

BBC (2017) Cyber-attack: europol says it was unprecedented in scale—BBC News. http://www.bbc.com/news/world-europe-39907965. Accessed 17 Jul 2017

Bender J (2014) Russia may still have an automated nuclear launch system aimed across the northern hemisphere. In: Bus. Insid. https://www.businessinsider.com.au/russias-dead-hand-system-may-still-be-active-2014-9. Accessed 17 Jul 2017

Blair BG (2011) The logic of accidental nuclear war. Brookings Institution Press, Washington, DC

Boles KS, Kannan K, Gill J et al (2017) Digital-to-biological converter for on-demand production of biologics. Nat Biotechnol 35:672–675 2017

Bostrom N (2002) Existential risks: analyzing human extinction scenarios and related hazards. J Evol Technol 9(1):1–30

Bostrom N (2003a) Astronomical waste: The opportunity cost of delayed technological development. Utilitas 15:308–314

Bostrom N (2003b) Are you living in a computer simulation? Publ Philos Q 53(211):243–255

Bostrom N (2006) What is a singleton. Linguist Philos Investig 5:48–54

Bostrom N (2009) Pascal's mugging. Analysis 69(3):443–445

Bostrom N (2011) Infinite ethics. Anal Metaphys 9–59

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

Bradbury RJ (2001) Matrioshka brains. preprint. http://www.aeiveos.com/bradbury/MatrioshkaBrains/MatrioshkaBrains.html

Carrigan RA Jr (2006) Do potential SETI signals need to be decontaminated? Acta Astronaut 58:112–117

Chalmers DJ (2002) Does conceivability entail possibility? In: Gendler T, Hawthorne J (eds) Conceivability possibility. Oxford University Press, New York pp 145–200

Chiew KL, Yong KSC, Tan CL (2018) A survey of phishing attacks: their types, vectors and technical approaches. Expert Syst Appl 106:1–20

Christiano P (2016) Prosaic AI alignment. https://ai-alignment.com/prosaic-ai-control-b959644d79c2. Accessed 27 Apr 2018
Clavero M, García-Berthou E (2005) Invasive species are a leading cause of animal extinctions. Trends Ecol Evol 20:110
Cole DD, Denkenberger D, Griswold M et al (2016) Feeding everyone if industry is disabled. In: Proceedings of the 6th international disaster and risk conference. Davos, Switzerland
Critch A (2017) Toward negotiable reinforcement learning: shifting priorities in Pareto optimal sequential decision-making (**arXiv:1701.01302**)
Daniel M (2017) S-risks: why they are the worst existential risks, and how to prevent them (EAG Boston 2017). https://foundational-research.org/s-risks-talk-eag-boston-2017/. Accessed 27 Apr 2018
Dennett DC (1978) Why you can't make a computer that feels pain. Synthese 38:415–456
Ellison H (1967) I have no mouth, and i must scream. Galaxy Publishing Corp, New York
Enserink M (2011) Scientists brace for media storm around controversial flu studies. In: Sciencemag. http://www.sciencemag.org/news/2011/11/scientists-brace-media-storm-around-controversial-flu-studies. Accessed 27 Apr 2018
Freitas R (2000) Some limits to global ecophagy by biovorous nanoreplicators, with public policy recommendations. Foresight Institute Technical Report
Future of Life Institute (2016) Accidental nuclear war: a timeline of close calls. https://futureoflife.org/background/nuclear-close-calls-a-timeline/. Accessed 4 Nov 2017
Futureworld (2013) Airplane "crashes" as hacker gets control. In: Futureworld. http://www.futureworld.org/PublicZone/MindBullets/MindBulletsDetails.aspx?MindBulletID=498. Accessed 27 Apr 2018
Gildert S (2011) Why "computronium" is really "unobtanium" IO9. http://io9.gizmodo.com/5758349/why-computronium-is-really-unobtanium. Accessed 27 Apr 2018
Goertzel B (2012) Should humanity build a global ai nanny to delay the singularity until it's better understood? J Conscious Stud 19(1–2):96–111
Grace K, Salvatier J, Dafoe A et al (2017) When will AI exceed human performance? evidence from AI experts. (**arXiv:1705.08807 [cs.AI]**)
Granoff J (2016) Donald trump is an existential threat to America and the world. Time
Gwern (2016) Why tool AIs want to be agent AIs. https://www.gwern.net/Tool-AI
Hanson R (2008) Catastrophe, social collapse, and human extinction. In: Bostrom N, Cirkovic MM (eds) Global catastrophic risks. Oxford University Press, Oxford, p 554
Hanson R (2016) The age of Em: work, love, and life when robots rule the earth. Oxford University Press, Oxford
Hines N (2016) Neural implants could let hackers hijack your brain. In: Inverse. https://www.inverse.com/article/19148-neural-implants-could-let-hackers-hijack-your-brain. Accessed 17 Jul 2017
Hume D (1739) A treatise of human nature. Oxford: Clarendon Press, London, UK
Hutter M (2000) A theory of universal artificial intelligence based on algorithmic complexity. ArXiv Prepr Cs0004001
Jenkins A (2018) Uber may not be to blame for self-driving car death in Arizona. Fortune, New York
Joy B (2000) Why the future doesn't need us. Wired, San Francisco, CA
Kahn H (1959) On thermonuclear war. Princeton University Press, Princeton
Kardashev NS (1985) On the inevitability and the possible structures of supercivilizations. Reidel Publishing Co., Dordrecht, pp 497–504
Karpathy A (2015) The unreasonable effectiveness of recurrent neural networks. Andrej Karpathy Blog. http://karpathy.github.io/2015/05/21/rnn-effectiveness/
Kushner D (2013) The real story of stuxnet. IEEE Spectr 50:48–53
LaVictoire P, Fallenstein B, Yudkowsky E et al (2014) Program equilibrium in the prisoner's dilemma via Löb's theorem. MIRI
LaVictorie P (2015) An Introduction to Löb's Theorem in MIRI Research. MIRI, San Francisco CA. http://intelligence.org/files/lob-notes-IAFF.pdf
Lem S (1961) Return from the stars. Houghton Mifflin Harcourt, Boston, US
Lem S (1963) Summa technologiae. Suhrkamp, Berlin, Germany
Lem S (1973) The Invincible: science fiction. Sidgwick & Jackson, London, UK
Lenat DB, Brown JS (1984) Why AM and EURISKO appear to work. Artif Intell 23:269–294
LoPucki LM (2017) Algorithmic ENTITIES. Social Science Research Network, Rochester
Lubin G (2016) Data reveals the 20 most popular TV shows of 2016. Business Insider
Mennen A (2017) Existential risk from AI without an intelligence explosion. http://lesswrong.com/lw/p28/existential_risk_from_ai_without_an_intelligence/
Menzel C (2017) Actualism. In: Zalta EN (ed) The stanford encyclopedia of philosophy, 2014th edn. Metaphysics Research Lab, Stanford University, Stanford
Meuhlhauser L (2014) How big is the field of artificial intelligence? (initial findings). https://intelligence.org/2014/01/28/how-big-is-ai/. Accessed 27 Apr 2018
Muehlhauser L (2011) Intelligence explosion FAQ. https://intelligence.org/ie-faq/. Accessed 27 Apr 2018
Mullin G (2017) What is the Blue Whale suicide challenge, how many deaths has the game been linked to so far and is it in the UK? TheSun
Oberhaus D (2017) Watch 'Slaughterbots', a warning about the future of killer bots. In: Motherboard. https://motherboard.vice.com/en_us/article/9kqmy5/slaughterbots-autonomous-weapons-future-of-life. Accessed 17 Dec 2017
Omohundro S (2008) The basic AI drives. In: Wang P, Goertzel B, Franklin S (eds) Proceedings of the 2008 conference on Artificial General Intelligence 2008: proceedings of the First AGI Conference. IOS Press Amsterdam, The Netherlands
Orwell G (1948) 1984. Houghton Mifflin Harcourt, Boston, US
Pinker S (2011) The better angels of our nature: The decline of violence in history and its causes. Penguin, London
Reason J (2000) Human error: models and management. BMJ 320:768–770
Russell S (2017) 3 principles for creating safer AI. https://www.youtube.com/watch?v=EBK-a94IFHY. Accessed 27 Apr 2018
Saito T, Angles J (2013) Hikikomori: adolescence without end. Univesity Of Minnesota Press, Minnesota
Sarma GP, Hay NJ (2016) Mammalian value systems. **(arXiv:1607.08289 [cs.AI])**
Schneier B (2017) Perspective | The next ransomware attack will be worse than WannaCry. Wash, Post
Shakirov V (2016) Review of state-of-the-arts in artificial intelligence with application to AI safety problem. **(ArXiv Prepr ArXiv160504232)**
Shulman C (2010) Omohundro's "basic AI drives" and catastrophic risks. http://intelligence.org/files/BasicAIDrives.pdf. Accessed 27 Apr 2018
Shulman C (2011) Arms races and intelligence explosions. Singularity Hypotheses. Springer, New York
Sotala K (2016) Decisive strategic advantage without a hard takeoff. http://kajsotala.fi/2016/04/decisive-strategic-advantage-without-a-hard-takeoff/#comments. Accessed 27 Apr 2018

Sotala K (2017) Disjunctive AI scenarios: Individual or collective takeoff? http://kajsotala.fi/2017/01/disjunctive-ai-scenarios-individual-or-collective-takeoff/. Accessed 27 Apr 2018
Sotala K, Yampolskiy R (2014) Responses to catastrophic AGI risk: a survey. Phys Scr 90:018001
Srugatsky N, Strugatsky B (1985) The time wanderers. Richardson & Steirman, New York, US
Strugatsky A, Strugatsky B (1976) The final circle of paradise, Translated by Leonid Renen. DAW, New York
Taylor A (2017) Flying around the world in a solar powered plane—the Atlantic
The Telegraph (2009) Russian spacecraft landed on moon hours before Americans. The telegraph. http://www.telegraph.co.uk:80/science/space/5737854/Russian-spacecraft-landed-on-moon-hours-before-Americans.html. Accessed 27 Apr 2018
Torres P (2014) Why running simulations may mean the end is near. https://ieet.org/index.php/IEET2/more/torres20141103. Accessed 27 Apr 2018
Torres P (2016) Problems with defining an existential risk. IEET. https://ieet.org/index.php/IEET2/more/torres20150121. Accessed 27 Apr 2018
Turchin A (2018) The risks connected with possibility of finding alien AI code during SETI. Rev J Br Interplanet Soc. Manuscript, https://philpapers.org/rec/TURCSW
Turchin A, Denkenberger D (2017) Levels of self-improvement. Manuscript, University of Louisville, TN
Turchin A, Denkenberger D (2018a) Military AI as convergent goal of the self-improving AI. In: Yampolskiy R (ed) Artificial intelligence safety and security. CRC Press, Baca Raton
Turchin A, Denkenberger D (2018b) Could slaughterbots wipe out humanity? Assessment of the global catastrophic risk posed by autonomous weapons. Manuscript
Turchin A, Green B, Denkenberger D (2017) multiple simultaneous pandemics as most dangerous global catastrophic risk connected with bioweapons and synthetic biology. Rev Health Secur
Turing AM (1937) On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 2:230–265
Velicovich B (2017) I could kill you with a consumer drone. Defense one, Washington, DC
Watkins J (2016) "Shut up and dance"—"Black mirror" series
Wei D (2013) Outside view(s) and MIRI's FAI endgame. http://lesswrong.com/lw/ig9/outside_views_and_miris_fai_endgame/. Accessed 27 Apr 2018
Wootson J (2017) Elon Musk doesn't think we're prepared to face humanity's biggest threat: artificial intelligence. Wash, Post
Yampolskiy R (2014) Utility function security in artificially intelligent agents. J Exp Theor Artif Intell JETAI 373–389. https://doi.org/10.1080/0952813X.2014.895114
Yampolskiy R (2015a) Artificial superintelligence: a futuristic approach. CRC Press, Boca Raton
Yampolskiy R (2015b) Taxonomy of pathways to dangerous AI. **(ArXiv Prepr ArXiv151103246)**
Yampolskiy R, Spellchecker M (2016) artificial intelligence safety and cybersecurity: a timeline of AI failures. **(ArXiv Prepr ArXiv161007997)**
Yudkowsky E (2001) Creating friendly AI 1.0: the analysis and design of benevolent goal architectures. MIRI, San Francisco, CA, pp 1–282
Yudkowsky E (2002) The AI-Box Experiment. http://yudkowsky.net/singularity/aibox. Accessed 27 Apr 2018
Yudkowsky E (2003) HUMOR: friendly AI critical failure table. http://www.sl4.org/archive/0310/7163.html. Accessed 27 Apr 2018
Yudkowsky E (2004) Coherent extrapolated volition. http://intelligence.org/files/CEV.pdf. Accessed 27 Apr 2018
Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk, in global catastrophic risks. Oxford University Press, Oxford
Yudkowsky E (2015) From AI to zombies. MIRI, San Francisco, CA
Yudkowsky E (2017) Comment on paper clip maximiser scenario. http://www.jefftk.com/p/examples-of-superintelligence-risk#fb-886930452142_886983450932. Accessed 27 Apr 2018
Yudkowsky E, Hanson R (2008) The Hanson-Yudkowsky AI-foom debate. In: MIRI Technical report
Yudkowsky E, Herreshoff M (2013) Tiling agents for self-modifying AI, and the Löbian obstacle. Early Draft MIRI