What to Do with the Singularity Paradox?

Roman V. Yampolskiy

Abstract. The paper begins with an introduction of the Singularity Paradox, an observation that: "Superintelligent machines are feared to be too dumb to possess commonsense". Ideas from leading researchers in the fields of philosophy, mathematics, economics, computer science and robotics regarding the ways to address said paradox are reviewed and evaluated. Suggestions are made regarding the best way to handle the Singularity Paradox.

Keywords: AI-Box, Friendliness, Machine Ethics, Singularity Paradox.

1 Introduction to the Singularity Paradox

Many philosophers, futurologists and artificial intelligence researchers [55, 9, 75, 37, 45, 69, 2, 66] have conjectured that in the next 20 to 200 years a machine capable of at least human level performance on all tasks will be developed. Since such a machine would among other things be capable of designing the next generation of even smarter intelligent machines it is generally assumed that an intelligence explosion will take place shortly after such a technological self-improvement cycle begins [30]. While specific predictions regarding the consequences of such an intelligence singularity are varied from potential economic hardship [35] to the complete extinction of the humankind [69, 9], many of the involved researchers agree that the issue is of utmost importance and needs to be seriously addressed [15].

Investigators concerned with the existential risks posed to humankind by the appearance of superintelligence often describe what we shall call a *Singularity Paradox* (SP) as their main reason for thinking that humanity might be in danger. Briefly SP could be described as: *"Superintelligent machines are feared to be too dumb to possess commonsense."*

Roman V. Yampolskiy
Department of Computer Engineering and Computer Science
University of Louisville
e-mail: roman.yampolskiy@louisville.edu

V.C. Müller (Ed.): Philosophy and Theory of Artificial Intelligence, SAPERE 5, pp. 397–413.
springerlink.com

SP is easy to understand via some commonly cited examples. Suppose that scientists succeed in creating a superintelligent machine and order it to "make all people happy". Complete happiness for humankind is certainly a noble and worthwhile goal, but perhaps we are not considering some unintended consequences of giving such an order. Any human immediately understands what is meant by this request; a non-exhaustive list may include making all people healthy, wealthy, beautiful, talented, giving them loving relationships and novel entertainment. However, many alternative ways of "making all people happy" could be derived by a superintelligent machine. For example:

- Killing all people trivially satisfies this request as with 0 people around all of them are happy.
- Forced lobotomies for every man, woman and child might also accomplish the same goal.
- A simple observation that happy people tend to smile may lead to forced plastic surgeries to affix permanent smiles to all human faces.
- A daily cocktail of cocaine, methamphetamine, methylphenidate, nicotine, and 3,4-methylenedioxymethamph-etamine, better known as Ecstasy, may do the trick.

An infinite number of other approaches to accomplish universal human happiness could be derived. For a superintelligence the question is simply which one is fastest/cheapest (in terms of computational resources) to implement. Such a machine clearly lacks commonsense, hence the paradox.

2 Methods Proposed for Dealing with SP

Prevention of Development

One of the earliest and most radical critics of the upcoming singularity was Theodore Kaczynski, a Harvard educated mathematician also known as the Unabomber. His solution to preventing singularity from ever happening was a bloody multiyear terror campaign against university research labs across the USA. In his 1995 manifesto Kaczynski explains his negative views regarding future of humankind dominated by the machines [44]: "*First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better that human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. … If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines."*

An even more violent outcome is prophesized, but not advocated, by Hugo de Garis [21] who predicts that the issue of building superintelligent machines will split humanity into two camps, eventually resulting in a civil war over the future of singularity research: "I believe that the ideological disagreements between these two groups on this issue will be so strong, that a major … war, killing billions of people, will be almost inevitable before the end of the 21st century".

Realizing potential dangers of superintelligent computers Anthony Berglas proposed a legal solution to the problem. He suggested outlawing production of more powerful processors essentially stopping Moore's Law in its tracks and consequently denying necessary computational resources to self-improving artificially intelligent machines [7]. Similar laws aimed at promoting human safety have been passed banning research on cloning of human beings and development of biological (1972 Biological Weapons Convention), chemical (1993 Chemical Weapons Convention) and nuclear weaponry. The idea of Berglas may be interesting in terms of its shock value which in turn may attract more attention to the dangers of the Singularity Paradox. Here is what Berglas suggested in his own words [7]: "… a radical solution, namely to limit the production of ever more powerful computers and so try to starve any AI of processing power. This is urgent, as computers are already almost powerful enough to host an artificial intelligence. … One major problem is that we may already have sufficient power in general purpose computers to support intelligence. Particularly if processors are combined into super computers or botnets. … So ideally we would try to reduce the power of new processors and destroy existing ones."

Alternatively restrictions could be placed on the intelligence an AI may possess to prevent it from becoming superintelligent [25] or legally require that its memory be erased after every job [6]. Similarly, Bill Joy advocates for relinquishment of superintelligence research and even suggests how enforcement of such convention could be implemented [43]: "*… enforcing relinquishment will require a verification regime similar to that for biological weapons, but on an unprecedented scale."* Enforcement of such technology restricting laws will not be trivial unless the society as a whole adopts an Amish-like, technology free, life style.

Ben Goertzel, a computer scientist, has proposed creation of "Big Brother AI" monitoring system he calls the "Singularity Steward". The goal of the proposed system is to monitor the whole world with the specific aim of preventing development of any technology capable of posing a risk to humanity including superintelligent machines [28]. Goertzel believes that creation of such a system is feasible and would safeguard humanity against preventable existential risks.

2.1 Restricted Deployment

A common theme in singularity discussion forums is a possibility of simply keeping a superintelligent agent in a sealed hardware so as to prevent it from doing any harm to the humankind [68]. Such ideas originate with scientific visionaries such as Eric Drexler who has suggested confining transhuman machines so that their outputs could be studied and used safely [18]. The general consensus on such an approach among researchers seems to be that such confinement is impossible to successfully maintain. For example, Vernor Vinge has strongly argued against the case of physical confinement [60]: *"Imagine yourself locked in your home with only limited data access to the outside, to your masters. If those masters thought at a rate − say −− one million times slower than you, there is little doubt that over a period of years (your time) you could come up with "helpful advice" that would incidentally set you free. (I call this "fast thinking" form of superintelligence*

"weak superhumanity". Such a "weakly superhuman" entity would probably burn out in a few weeks of outside time. "Strong superhumanity" would be more than cranking up the clock speed on a human−equivalent mind. It's hard to say precisely what "strong superhumanity" would be like, but the difference appears to be profound."

Likewise David Chalmers, a philosopher, has stated that confinement is impossible as any useful information we would be able to extract from the AI will affect us, defeating the purpose of confinement [15]. However, the researcher who did the most to discredit the idea of the so called "AI-Box" is Eliezer Yudkowsky who has actually performed AI-Box "experiments" in which he demonstrated that even human level intelligence is sufficient to escape from an AI-Box [71]. In a series of 5 experiments, Yudkowsky has challenged different individuals to play a role of a gatekeeper to a Superintelligent Agent (played by Yudkowsky himself) trapped inside an AI-Box, and was successful in securing his release in 3 out of 5 trials via nothing more than a chat interface [71].

In 2010 David Chalmers proposed the idea of a "leakproof" singularity. He suggests that for safety reasons, first AI systems be restricted to simulated virtual worlds until their behavioral tendencies could be fully understood under the controlled conditions. Chalmers argues that even if such an approach is not foolproof, it is certainly safer than building AI in physically embodied form. However, he also correctly observes that a truly leakproof system in which no information is allowed to leak out from the simulated world into our environment "… is impossible, or at least pointless" [15] since we can't interact with the system or even observe it. Chalmers' discussion of the leakproof singularity is an excellent introduction to the state-of-the-art thinking in the field.

Nick Bostrom, a futurologist, has proposed [10] an idea for an Oracle AI (OAI), which would be only capable of answering questions. It is easy to elaborate and see that a range of different Oracle AIs is possible. From advanced OAIs capable of answering any question to domain-expert-AIs capable of answering Yes/No/Unknown to questions on a specific topic. It is claimed that an OAI could be used to help mankind build a safe unrestricted superintelligent machine.

2.2 Incorporation into Society

Robin Hanson has suggested that as long as future intelligent machines are law abiding they should be able to coexist with humans [36]. Similarly, Hans Moravec puts his hopes for humanity in the hands of the law. He sees forcing cooperation from the robot industries as the most important security guarantee for humankind, and integrates legal and economic measures into his solution [43]. Robin Hanson, an economist, agrees [35]: "…robots well-integrated into our economy would be unlikely to exterminate us." Similarly, Steve Omohundro uses micro-economic theory to speculate about the driving forces in the behavior of superintelligent machines. He argues that intelligent machines will want to self-improve, be rational, preserve their utility functions, prevent counterfeit utility, acquire resources and use them efficiently, and protect themselves. He believes that machines' actions will be governed by rational economic behavior [50, 49].

Mark Waser suggested an additional "drive" to be included in the list of behaviors predicted to be exhibited by the machines [63]. Namely, he suggests that evolved desires for cooperation and being social are part of human ethics and are a great way of accomplishing goals, an idea also analyzed by Fox et al., who come to the conclusion that superintelligence does not imply benevolence [19]. Bill Hibbard adds the desire for maintaining the social contract towards equality as a component of ethics for super-intelligent machines [40] and J. Storrs Hall argues for incorporation of moral codes into the design [34]. In general ethics for superintelligent machines is one of the most fruitful areas of research in the field of singularity research with numerous publications appearing every year [53, 11, 9, 56, 54, 62, 13].

Robert Geraci, a theologian, has researched similarities between different aspects of technological singularity and the world's religions [24]. In particular, in his work on Apocalyptic AI [22] he observes the many commonalities in the works of Biblical prophets like Isaiah and the prophets of the upcoming technological singularity such as Ray Kurzweil or Hans Moravec. All promise freedom from disease, immortality, and purely spiritual (software) existence in the Kingdom come (Virtual Reality). More interestingly Geraci argues [23] that in order to be accepted into the society as equals, robots must convince most people that they are conscious beings. Geraci believes that an important component for such attribution is voluntary religious belief. Just like some people choose to believe in a certain religion, so will some robots. In fact one may argue that religious values may serve the goal of limiting behavior of superintelligences to those acceptable to society just like they do for many people. David Brin, in a work of fiction, has proposed that smart machines should be given humanoid bodies and from inception raised as our children and taught the same way we were [12]. Instead of programming machines explicitly to follow a certain set of rules they should be given capacity to learn and should be immersed in human society with its rich ethical and cultural rules.

2.3 Self-Monitoring

Probably the earliest and the best known solution for the problem of intelligent machines has been proposed by Isaac Asimov, a biochemist and a science fiction writer, in the early 1940s. The so called "Three Laws" of robotics are almost universally known and have inspired numerous imitations as well as heavy critique [32, 47, 65, 51]. The original laws as given by Asimov are [4]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with either the First or Second Law.

Continuing Asimov's work, rule-based standards of behavior for robots have been recently proposed by South Korea's Ministry of Commerce, Industry, and Energy. In 2007 a Robot Ethics Charter, which sets ethical guidelines concerning robot

functions has been adopted. In Europe, EURON (the European Robotics Research Network) also announced plans to develop guidelines for robots in five areas: safety, security, privacy, traceability, and identifiability. Japan's Ministry of Economy, Trade, and Industry has issued policies regarding robots in homes and how they should behave and be treated [52].

Stuart Armstrong proposed that trustworthiness of a superintelligent system could be monitored via a chain of progressively less powerful AI systems all the way down to the human level of intelligence [3]. The proposed "chain" would allow people to indirectly monitor and perhaps control the ultraintelligent machine. However, Armstrong himself acknowledges a number of limitations of the proposed method: the meaning of communication could be lost from one AI level to the next or AI links in the chain may not be able to reliably judge the trustworthiness of a more intelligent entity. In such cases the proposed solution is to shut down all AIs and to start building the chain from scratch.

To protect humankind against unintended consequences of superintelligent machines Eliezer Yudkowsky, an AI researcher, has suggested that any AI system under development should be "Friendly" to humanity [69]. Friendliness according to Yudkowsky could be defined as looking out for the best interests of the humankind. To figure out what humankind is really interested in, design of Friendly AI (FAI) should be done by specialized AIs. Such Seed AI [74] systems will first study human nature and then produce a Friendly Superintelligence humanity would want if it was given sufficient time and intelligence to arrive at a satisfactory design, our Coherent Extrapolated Volition (CEV) [72]. Yudkowsky is not the only researcher working on the problem of extracting and understanding human desires, Tim Freeman has also attempted to formalize a system capable of such "wish-mining" but in the context of "compassionate" and "respectful" plan development by AI systems [20].

For Friendly self-improving AI systems a desire to pass friendliness as a main value to the next generation of intelligent machines should be a fundamental drive. Yudkowsky also emphasizes importance of the "first mover advantage" - the first superintelligent AI system will be powerful enough to prevent any other AI systems from emerging, which might protect humanity from harmful AIs. Here is how Yudkowsky himself explains FAI [73] and CEV [72]: *"The term "Friendly AI" refers to the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals."* *"… our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted."*

Ben Goertzel, a frequent critic of Friendly AI [27] has proposed a variation on the theme he calls a Humane AI. He believes it is more feasible to install AI with general properties like compassion, choice and growth than with specific properties like friendliness to humans [27]. In Goertzel's own words [28]: "In Humane AI, one posits as a goal, not simply the development of AI's that are benevolent to

humans, but the development of AI's that display the qualities of "humaneness," … That is, one proposes "humaneness" as a kind of ethical principle, where the principle is: "Accept an ethical system to the extent that it agrees with the body of patterns known as 'humaneness'."

Bill Hibbard believes that the design of superintelligent machines needs to incorporate emotions that can guide the process of learning and self-improvement in such machines. In his opinion machines should love us as their most fundamental emotion and consequently they will attempt to make us happy and prosperous. He states [41]: "So in place of laws constraining the behavior of intelligent machines, we need to give them emotions that can guide their learning of behaviors." Others have also argued for importance of emotions, for example Mark Waser wrote [63]: "…thinking machines need to have analogues to emotions like fear and outrage that create global biases towards certain actions and reflexes under appropriate circumstances".

2.4 Indirect Solutions

Continuing with the economic model of supply and demand it is possible to argue that the superintelligent machines will need humans and therefore not exterminate humanity (but still might treat it less than desirably). For example in the movie Matrix, machines need the heat from our bodies as energy. It is not obvious from the movie why this would be an efficient source of energy but we can certainly think of other examples.

Friendly AI is attempting to replicate what people would refer to as "common sense" in the domain of plan formation [70]. Since only humans know what it is like to be a human [48] the Friendly machines would need people to provide that knowledge, to essentially answer the question: "What Would Human Do (WWHD)?"

Alan Turing in "*Intelligent Machinery, a Heretical Theory*" argued that humans can do something machines can't, namely overcome limitations of Godel's incompleteness theorem [58]. Here is what Turing said on this matter [58]: "*By Godel's famous theorem, or some similar argument, one can show that however the machine is constructed there are bound to be cases where the machine fails to give an answer, but a mathematician would be able to.*"

Another area of potential need for assistance from human beings for machines may be deduced from some peer-reviewed experiments showing that human consciousness can affect Random Number Generators and other physical processes [5]. Perhaps ultraintelligent machines will want that type of control or some more advanced technology derivable from it.

As early as 1863 Samuel Butler has argued that the machines will need us to help them reproduce: "*They cannot kill us and eat us as we do sheep; they will not only require our services in the parturition of their young (which branch of their economy will remain always in our hands), but also in feeding them, in setting them right when they are sick, and burying their dead or working up their corpses into new machines."* [14].

A set of anthropomorphic arguments is also often made. They usually go something like: by analyzing human behavior we can see some reasons for a particular type of intelligent agent not to exterminate a less intelligent life form. For example, humankind doesn't need elephants and we are smarter and certainly capable of wiping them out but instead we spend lots of money and energy on preserving them, why? Is there something inherently valuable in all life forms? Perhaps their DNA is great source of knowledge which we may later use to develop novel medical treatments? Or maybe their minds could teach us something? Maybe the fundamental rule implanted in all intelligent agents should be that information should never be destroyed. As each living being is certainly packed with unique information this would serve as a great guiding principle in all decision making. Similar arguments could be made about the need of superintelligent machines to have cute human pets, or a desire for companionship with other intelligent species, or a milliard other human needs. For example, Mark Waser, a proponent of teaching the machines universal ethics [64], which only exist in the context of society, suggested that we should "… convince our super-intelligent AIs that it is in their own self-interest to join ours."

Some scientists are willing to give up on humanity all together in the name of a greater good that they claim ultraintelligent machines will bring [17]. They see machines as the natural next step in evolution and believe that humanity has no right to stand in the way of progress. Essentially their position is - let the machines do what they want, they are the future, no humanity is not necessarily a bad thing. They may see desire to keep humanity alive as nothing but a self-centered bias of Homo sapiens. Some may even give reasons for why humanity is undesirable to nature such as environmental impact on Earth and later on maybe the cosmos at large. To quote from some of the proponents of the "let them kill us" philosophy: "*Humans should not stand in the way of a higher form of evolution. These machines are godlike. It is human destiny to create them"* [1] believes Hugo de Garis.

Amazingly as early as 1863 Samuel Butler has written about the need for a violent struggle against machine oppression: "*… the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question. Our opinion is that war to the death should be instantly proclaimed against them."* [14].

An alternative vision for the post singularity future of humanity could be summarized as: "If you can't beat them, join them". A number of prominent scientists have suggested pathways for humanity to be able to keep up with superintelligent machines by becoming partially or completely merged with our engineered progeny. Ray Kurzweil is an advocate of a process known as uploading in which a mind of a person is scanned and copied into a computer [45]. The specific pathway to such scanning is not important but suggested approaches include advanced Brain Computer Interfaces (BCI), brain scanning and nanobots. A copied human could either reside in robotic body or in virtual reality. In any case superior computational resources in terms of processing speed and memory become available to such an uploaded human making it feasible for the person to keep up with superintelligent machines.

A slightly less extreme approach is proposed by Kevin Warwick who also agrees that we will merge with our machines but via direct integration of our bodies with them. Devices such as brain implants will give "cyborgs" computational resources necessary to compete with the best of the machines. Novel sensors will provide sensual experiences beyond the five we are used to operating with. A human being with direct uplink to the wireless Internet will be able to instantaneously download necessary information or communicate with other cyborgs [61]. Both Kurzweil and Warwick attempt to analyze potential consequences of humanity joining the machines and come up with numerous fascinating predictions. The one aspect they agree on is that humanity will never be the same. Peter Turney suggests an interesting twist on the "fusion" scenario: "*One approach to controlling a [superintelligence] would be to link it directly to a human brain. If the link is strong enough, there is no issue of control. The brain and the computer are one entity; therefore, it makes no sense to ask who is controlling whom."* [59].

3 Other Approaches

While we have reviewed some of the most prominent and frequently suggested approaches for dealing with the Singularity Paradox many other approaches and philosophical viewpoints are theoretically possible. Many of them would fall into the Singularity "denialist" camp accepting the following statement by Jeff Hawkins [2]: "There will be no singularity or point in time where the technology itself runs away from us." He further elaborates [2]: "Exponential growth requires the exponential consumption of resources (matter, energy, and time), and there are always limits to this. Why should we think intelligent machines would be different? We will build machines that are more 'intelligent' than humans and this might happen quickly, but there will be no singularity, no runaway growth in intelligence. There will be no single godlike intelligent machine." A recent report from the AAAI presidential panel on long-term AI futures outlines similar beliefs held by the majority of the participating AI scientists: "There was overall skepticism about the prospect of an intelligence explosion as well as of a "coming singularity," and also about the large-scale loss of control of intelligent systems" [42].

Others may believe that we might get lucky and even if we do nothing the superintelligence will turn out to be friendly to us and possess some human characteristics. Perhaps this will happen as a side effect of being (directly or indirectly) designed by human engineers who will, maybe subconsciously, incorporate such values into their designs or as Douglas Hofstadter put it [2]: "Perhaps these machines--our 'children'--will be vaguely like us and will have culture similar to ours…". Yet others think that superintelligent machines will be neutral towards us. John Casti thinks that [2]: "… machines will become increasingly uninterested in human affairs just as we are uninterested in the affairs of ants or bees. But it's more likely than not in my view that the two species will comfortably and more or less peacefully coexist…". Both Peter Turney [59] and Alan Turing [57] suggested that giving machines an ability to feel pleasure and pain will allow us to control them to a certain degree and will assist in machine learning. Unfortunately teaching machines to feel pain is not an easy problem to solve [8, 16]. Finally, one

can simply deny that the problem exists by questioning either possibility of the technological singularity or not accepting that it leads to the Singularity Paradox. Perhaps one can believe that a superintelligent machine by its very definition will have at least as much common sense as an average human and will consequently act accordingly.

4 Analysis of Solutions

In this paper we provide an overview of methods which were proposed to either directly or indirectly address the problem we have named the Singularity Paradox. We have categorized the proposed solutions into five broad categories, namely: Prevention of Development, Restricted Deployment, Incorporation into Society, Self-Monitoring, and Indirect Solutions. Such grouping makes it easier to both understand the proposed methods and to analyze them as a set of complete measures. We will review each category and analyze it in terms of feasibility of accomplishing the proposed actions and more importantly try to evaluate the likelihood of the method succeeding if implemented.

Violent struggle against scientific establishment, outlawing AI research and placing restrictions on development and sale of hardware components are all a part of an effort to prevent superintelligent machines from ever coming into existence and to some extent are associated with the modern Luddite movement. Given the current political climate, complex legal system and economic needs of the world's most developed countries it is highly unlikely that laws will be passed to either ban computer scientists from researching AI systems or from developing and selling faster processors. Since for this methodology to work the ban needs to be both global and enforceable it will not work as there is no global government to enforce such a law or to pass it in the first place. Even if such a law was passed there is always a possibility that some rogue scientist somewhere will simply violate the restrictions making it at best a short term solution.

An idea for an automated monitoring system AKA "Big Brother AI" is as likely to be accepted by humanity as the legal solution analyzed above. It also presents the additional challenge of technological implementation which as far as we can tell would be as hard to make "humanity safe" as a full blown singularity level AI system. Provided that the system would have to be given legal rights to control people we can quote Martha Moody by saying "Sometimes the cure is worse than the disease." Finally, as for the idea of violent struggle, it may come to be, as suggested by Hugo de Garis [21] but we will certainly not advocate such an approach or even consider it as a real solution.

Restricting access of superintelligent machines to the real world is a commonly proposed solution to the SP problem. AI-boxes, Leakproofing and restricted question-answering-only systems known as Oracle AIs are just some of the proposed methods for accomplishing that. While a lot of skepticism has been expressed towards the possibility of long term restriction of a superintelligent mind no one so far has proven that it is impossible with mathematical certainty. This approach may be similar to putting a dangerous human being in prison. While some have escaped from even maximum security facilities, in general, prisons do provide a

certain measure of security which while not perfect is still beneficial for improving overall safety of the society. This approach may provide some short term relief especially in the early stages of the development of truly intelligent machines. We also feel that this area is one of the most likely to be accepted by the general scientific community as research in the related fields of computer and network security, steganography detection, computer viruses, encryption, and cyber-warfare is well funded and highly publishable. While without a doubt the restriction methodology will be extremely difficult to implement, it might serve as a tool for at least providing humanity with a little more time to prepare a better response.

Numerous suggestions for regulating behavior of machines by incorporating them into the human society have been proposed. Economic theories, legal recourse, human education, ethical principles of morality and equality, and even religious indoctrination have been suggested as a way to make superintelligent machines a part of our civilization. It seems that the proposed methods are a result of an anthropomorphic bias as it is not obvious why would machines with minds drastically different from human and which have no legal status, no financial responsibilities, no moral compass and no spiritual desires be interested in any of the typical human endeavors of daily life. We could of course try and program into the superintelligent machines such tendencies as meta-rules but then we simply change our approach to the so called "Self-Monitoring" methods which we will discuss later. While the ideas proposed in this category are straightforward to implement we are skeptical of their usefulness as any even slightly intelligent machine will discover all the loopholes in our legal, economic and ethical system as well or better as human beings are known to be able to. With respect to the idea of raising machines as our children and giving them a human education this would not only be impractical because of the required time but also because we all know about children who greatly disappoint their parents.

The Self-Monitoring category groups together very dissimilar approaches such as explicitly hard-coding rules of behavior into the machine, creating numerous levels of machines with increasing capacity to monitor each other or providing machines with a fundamental and unmodifiable desire to be nice to humanity. The idea of providing explicit rules for robots to follow is the oldest approach surveyed in this paper and as such has received the most criticism over the years. The general consensus seems to be that no set of rules can ever capture every possible situation and that interaction of rules may lead to unforeseen circumstances and undetectable loopholes leading to devastating consequences for humanity.

The approach of chaining multiple levels of AI systems with progressively greater capacity seems to be replacing a very difficult problem of solving SP with a much harder problem of solving a multi-system version of the same problem. Numerous issues with the chain could arise such as the break in the chain of communication or an inability of a system to accurately assess the mind of another (especially smarter) system. Also the process of constructing the chain is not trivial.

Finally the approach of making a fundamentally friendly system which will desire to preserve its friendliness under numerous self-improvement measures seems to be very likely to work if implemented correctly. Unfortunately no one knows

how to create a human-friendly self-improving optimization process and some have argued that it is impossible [46, 29, 26]. It is also unlikely that creating a friendly intelligent machine is easier than creating any intelligent machine, creation of which would still produce a Singularity Paradox. Similar criticism could be applied to many variations on the Friendly AI theme for example Goertzel's Humane AI or Freeman's Compassionate AI. As one of the more popular solutions to the SP problem the Friendliness approach has received a significant dose of criticisms [27, 39, 38], however we believe that this area of research is well suited for scientific investigation and further research by the main stream AI community. Work has already begun in the general area of assuring the behavior of intelligent agents [31, 33].

To summarize our analysis of Self-Monitoring methods we can say that explicit rules are easy to implement, but are unlikely to serve the intended purpose. The chaining approach is too complex to implement or verify and has not been proven to be workable in practice. Finally, the approach of installing fundamental desire into the superintelligent machines to treat humanity nicely may work if implemented but as of today no one can accurately evaluate feasibility of such an implementation. Finally, the category of Indirect Approaches is comprised of nine highly diverse methods some of which are a bit extreme and others provide no solution at all. For example Peter Turney's idea of giving machines the ability to feel pleasure and pain does not in any way prevent machines from causing humanity great amounts of the latter and in fact may help machines in becoming torture experts given their personal experiences with pain.

The next approach is based on the idea first presented by Samuel Butler and later championed by Alan Turing and others, is that the machines will need us for some purpose, such as procreation, and so will treat us nicely. This is highly speculative and it requires us to prove existence of some property of human beings for which superintelligent machines will not be able to create a simulator (reproduction is definitely not such a property for software agents). This is highly unlikely and even if there is such a property it does not guarantee nice treatment of humanity, since just one of us may be sufficient to perform the duty or maybe even a dead human will be as useful in supplying the necessary degree of humanness.

A very extreme view is presented (at least in the role of Devil's advocate) by Hugo de Garis who says that the superintelligent machines are better than us and so deserve to take over even if it means the end of the human race. While it is certainly a valid philosophical position it is neither a solution to the SP nor a desirable outcome in the eyes of the majority of people. Likewise, Butler's idea of an outright war against superintelligent machines is likely to bring humanity to extinction due to the shear difference in capabilities between the two types of minds.

Another non-solution is discussed by Jeff Hawkins who simply states that the Technological Singularity will not happen and so consequently SP will not be a problem. Others admit that the Singularity may take place but think that we may get lucky and the machines will be nice to us just by chance. Neither one of those positions offers much in terms of solution and the chances of us getting lucky given the space of all possible non-human minds is very close to zero.

Finally, a number of hybrid approaches are suggested which say that instead of trying to control or defeat the superintelligent machines we should join them. Either via brain implants or via uploads we could become just as smart and powerful as machines, defeating the SP problem by supplying our common sense to the machines. In our opinion the presented solutions are both feasible (in particular the cyborgs option) to implement and is likely to work, unfortunately we may have a Pyrrhic victory. In the process of defending humanity we might lose ours. Last but not least, we have to keep in mind a possibility that the SP simply has no solution and prepare to face the unpredictable post-Singularity world.

5 Conclusions

With the survival of humanity on the line, the issues raised by the problem of the Singularity Paradox are too important to put "all our eggs in one basket". We should not limit our response to any one technique, or an idea from any one scientist or a group of scientists. A large research effort from the scientific community is needed to solve this issue of global importance [67]. Even if there is a relatively small chance that a particular method would succeed in preventing an existential catastrophe it should be explored as long as it is not likely to create significant additional dangers to the human race. After analyzing dozens of solutions from as many scientists, we came to the conclusion that the search is just beginning. Perhaps because the winning strategy has not yet been suggested or maybe additional research is needed to accept an existing solution with some degree of confidence.

For a long time work related to the issues raised in this volume has been informally made public via online forums, blogs and personal website by a few devoted enthusiasts. We believe the time has come for the singularity research to join mainstream science. It could be a field in its own right supported by strong interdisciplinary underpinnings and attracting top mathematicians, philosophers, engineers, psychologists, computer scientists and academics from other fields.

References

[1] Anonymous, Hugo de Garis, Wikipedia.org (1999), http://en.wikipedia.org/wiki/Hugo_de_Garis

[2] Anonymous, Tech Luminaries Address Singularity, IEEE Spectrum. Special Report: The Singularity (June 2008), http://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity

[3] Armstrong, S.: Chaining God: A qualitative approach to AI, trust and moral systems. New European Century (2007), http://www.neweuropeancentury.org/GodAI.pdf

[4] Asimov, I.: Runaround in Astounding Science Fiction (March 1942)

[5] Bancel, P., Nelson, R.: The GCP Event Experiment: Design, Analytical Methods, Results. Journal of Scientific Exploration 22(4) (2008)

[6] Benford, G.: "Me/Days", in Alien Flesh. Victor Gollancz, London (1988)

[7] Berglas, A.: Artificial Intelligence Will Kill Our Grandchildren (February 22, 2009), http://berglas.org/Articles/AIKillGrandchildren/AIKillGrandchildren.html
[8] Bishop, M.: Why Computers Can't Feel Pain. Minds and Machines 19(4), 507–516 (2009)
[9] Bostrom, N.: Ethical Issues in Advanced Artificial Intelligence. Review of Contemporary Philosophy 5, 66–73 (2006)
[10] Bostrom, N.: Oracle AI (2008), http://lesswrong.com/lw/qv/the_rhythm_of_disagreement/
[11] Bostrom, N., Yudkowsky, E.: The Ethics of Artificial Intelligence. In: Ramsey, W., Frankish, K. (eds.) Cambridge Handbook of Artificial Intelligence. Cambridge University Press (2011)
[12] Brin, D.: Lungfish (1987), http://www.davidbrin.com/lungfish1.html
[13] Bugaj, S., Goertzel, B.: Five Ethical Imperatives and their Implications for Human-AGI Interaction. Dynamical Psychology (2007), http://goertzel.org/dynapsyc/2007/Five_Ethical_Imperatives_svbedit.html
[14] Butler, S.: Darwin Among the Machines, To the Editor of Press, Christchurch, New Zealand, June 13 (1863)
[15] Chalmers, D.: The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17, 7–65 (2010)
[16] Dennett, D.C.: Why You Can't Make a Computer That Feels Pain. Synthese 38(3), 415–456 (1978)
[17] Dietrich, E.: After the Humans are Gone. Journal of Experimental & Theoretical Artificial Intelligence 19(1), 55–67 (2007)
[18] Drexler, E.: Engines of Creation. Anchor Press (1986)
[19] Fox, J., Shulman, C.: Superintelligence Does Not Imply Benevolence. In: 8th European Conference on Computing and Philosophy, Munich, Germany, October 4-6 (2010)
[20] Freeman, T.: Using Compassion and Respect to Motivate an Artificial Intelligence (2009), http://www.fungible.com/respect/paper.html
[21] Garis, H.D.: The Artilect War. ETC publications (2005)
[22] Geraci, R.M.: Apocalyptic AI: Religion and the Promise of Artificial Intelligence. The Journal of the American Academy of Religion 76(1), 138–166 (2008)
[23] Geraci, R.M.: Religion for the Robots, Sightings. Martin Marty Center at the University of Chicago, June 14 (2007), http://divinity.uchicago.edu/martycenter/publications/~sightings/archive_2007/0614.shtml
[24] Geraci, R.M.: Spiritual Robots: Religion and Our Scientific View of the Natural World. Theology and Science 4(3), 229–246 (2006)
[25] Gibson, W.: Neuromancer. Ace Science Fiction, New York (1984)
[26] Goertzel, B.: The All-Seeing (A)I. Dynamic Psychology (2004), http://www.goertzel.org/dynapsyc
[27] Goertzel, B.: Apparent Limitations on the "AI Friendliness" and Related Concepts Imposed By the Complexity of the World (September 2006), http://www.goertzel.org/papers/LimitationsOnFriendliness.pdf

[28] Goertzel, B.: Encouraging a Positive Transcension. Dynamical Psychology (2004), http://www.goertzel.org/dynapsyc/2004/PositiveTranscension.html
[29] Goertzel, B.: Thoughts on AI Morality. Dynamical Psychology (2002), http://www.goertzel.org/dynapsyc
[30] Good, I.J.: Speculations Concerning the First Ultraintelligent Machine. Advances in Computers 6, 31–88 (1966)
[31] Gordon-Spears, D.: Assuring the behavior of adaptive agents. In: Rouff, C.A., et al. (eds.) Agent Technology From a Formal Perspective, pp. 227–259. Kluwer (2004)
[32] Gordon-Spears, D.F.: Asimov's Laws: Current Progress. In: Hinchey, M.G., Rash, J.L., Truszkowski, W.F., Rouff, C.A., Gordon-Spears, D.F. (eds.) FAABS 2002. LNCS (LNAI), vol. 2699, pp. 257–259. Springer, Heidelberg (2003)
[33] Gordon, D.F.: Well-Behaved Borgs, Bolos, and Berserkers. In: 15th International Conference on Machine Learning (ICML 1998), San Francisco, CA (1998)
[34] Hall, J.S.: Ethics for Machines (2000), http://autogeny.org/ethics.html
[35] Hanson, R.: Economics of the Singularity. IEEE Spectrum 45(6), 45–50 (2008)
[36] Hanson, R.: Prefer Law to Values (October 10 , 2009), http://www.overcomingbias.com/2009/10/prefer-law-to-values.html
[37] Hawking, S.: Science in the Next Millennium. In: The Second Millennium Evening at The White House, Washington, DC, March 6 (1998)
[38] Hibbard, B.: Critique of the SIAI Collective Volition Theory (December 2005), http://www.ssec.wisc.edu/~billh/g/SIAI_CV_critique.html
[39] Hibbard, B.: Critique of the SIAI Guidelines on Friendly AI (2003), http://www.ssec.wisc.edu/~billh/g/SIAI_critique.html
[40] Hibbard, B.: The Ethics and Politics of Super-Intelligent Machines (July 2005), http://www.ssec.wisc.edu/~billh/g/SI_ethics_politics.doc
[41] Hibbard, B.: Super-Intelligent Machines. Computer Graphics 35(1), 11–13 (2001)
[42] Horvitz, E., Selman, B.: Interim Report from the AAAI Presidential Panel on Long-Term AI Futures (August 2009), http://aaai.org/Organization/Panel/panel-note.pdf
[43] Joy, B.: Why the Future Doesn't Need Us. Wired Magazine 8(4) (April 2000)
[44] Kaczynski, T.: Industrial Society and Its Future. The New York Times, September 19 (1995)
[45] Kurzweil, R.: The Singularity is Near: When Humans Transcend Biology. Viking (2005)
[46] Legg, S.: Friendly AI is Bunk, Vetta Project (2006), http://commonsenseatheism.com/wp-content/uploads/2011/02/
[47] Mccauley, L.: AI Armageddon and the Three Laws of Robotics. Ethics and Information Technology 9(2) (2007)
[48] Nagel, T.: What is it Like to be a Bat? The Philosophical Review LXXXIII(4), 435–450 (1974)
[49] Omohundro, S.M.: The Basic AI Drives. In: Wang, P., Goertzel, B., Franklin, S. (eds.) Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, vol. 171. IOS Press (February 2008)
[50] Omohundro, S.M.: The Nature of Self-Improving Artificial Intelligence, Singularity Summit, San Francisco, CA (2007)

[51] Pynadath, D.V., Tambe, M.: Revisiting Asimov's First Law: A Response to the Call to Arms. In: Meyer, J.-J.C., Tambe, M. (eds.) ATAL 2001. LNCS (LNAI), vol. 2333, p. 307. Springer, Heidelberg (2002)
[52] Sawyer, R.J.: Robot Ethics. Science 318, 1037 (2007)
[53] Shulman, C., Jonsson, H., Tarleton, N.: Machine Ethics and Superintelligence. In: 5th Asia-Pacific Computing & Philosophy Conference, Tokyo, Japan, October 1-2 (2009)
[54] Shuman, C., Tarleton, N., Jonsson, H.: Which Consequentialism? Machine Ehics and Moral Divergence. In: Asia-Pacific Conference on Computing and Philosophy (APCAP 2009), Tokyo, Japan, October 1-2 (2009)
[55] Solomonoff, R.J.: The Time Scale of Artificial Intelligence: Reflections on Social Effects. North-Holland Human Systems Management 5, 149–153 (1985)
[56] Sotala, K.: Evolved Altruism, Ethical Complexity, Anthropomorphic Trust. In: 7th European Conference on Computing and Philosophy (ECAP 2009), Barcelona, July 2-4 (2009)
[57] Turing, A.: Computing Machinery and Intelligence. Mind 59(236), 433–460 (1950)
[58] Turing, A.M.: Intelligent Machinery, A Heretical Theory. Philosophia Mathematica 4(3), 256–260 (1996)
[59] Turney, P.: Controlling Super-Intelligent Machines. Canadian Artif. Intell., 27 (1991)
[60] Vinge, V.: The Coming Technological Singularity: How to Survive in the Post-human Era. In: Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace, Cleveland, OH, March 30-31, pp. 11–22 (1993)
[61] Warwick, K.: Cyborg Morals, Cyborg Values, Cyborg Ethics. Ethics and Information Technology 5, 131–137 (2003)
[62] Waser, M.: Deriving a Safe Ethical Architecture for Intelligent Machines. In: 8th Conference on Computing and Philosophy (ECAP 2010), October 4-6 (2010)
[63] Waser, M.R.: Designing a Safe Motivational System for Intelligent Machines. In: The Third Conference on Artificial General Intelligence, Lugano, Switzerland, March 5-8 (2010)
[64] Waser, M.R.: Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence, AAAI Technical Report FS-08-04, Menlo Park, CA (2008)
[65] Weld, D.S., Etzioni, O.: The First Law of Robotics (a Call to Arms). In: National Conference on Artificial Intelligence, pp. 1042–1047 (1994)
[66] Yampolskiy, R.V.: AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System. ISRN Artificial Intelligence, 271878 (2011)
[67] [67] Yampolskiy, R.V.: Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach. In: Philosophy and Theory of Artificial Intelligence (PT-AI 2011), Thessaloniki, Greece, October 3-4 (2011)
[68] Yampolskiy, R.V.: Leakproofing Singularity - Artificial Intelligence Confinement Problem. Journal of Consciousness Studies (JCS), 19(1-2) (2012)
[69] Yudkowsky, E.: Artificial Intelligence as a Positive and Negative Factor in Global Risk. In: Bostrom, N., Cirkovic, M.M. (eds.) Global Catastrophic Risks, pp. 308–345. Oxford University Press, Oxford (2008)
[70] Yudkowsky, E.: What is Friendly AI? (2005), http://singinst.org/ourresearch/publications/what-is-friendly-ai.html

[71] Yudkowsky, E.S.: The AI-Box Experiment (2002), http://yudkowsky.net/singularity/aibox
[72] Yudkowsky, E.S.: Coherent Extrapolated Volition, Singularity Institute for Artificial Intelligence (May 2004), http://singinst.org/upload/CEV.html
[73] Yudkowsky, E.S.: Creating Friendly AI - The Analysis and Design of Benevolent Goal Architectures (2001), http://singinst.org/upload/CFAI.html
[74] Yudkowsky, E.S.: General Intelligence and Seed AI (2001), http://singinst.org/ourresearch/publications/GISAI/
[75] Yudkowsky, E.S.: Three Major Singularity Schools, Singularity Institute Blog (September 2007), http://yudkowsky.net/singularity/schools