AGI Risk and Friendly AI Policy Solutions

Chris Nota

March 6, 2015

Abstract

This paper introduces the risks associated with the development of Artificial General Intelligence (AGI), gives an overview of the major organizations involved in the study of AGI risk, and examines current actions being taken with respect to AGI risk management. It then discusses possible AGI public policy options and their likely outcomes, and finally recommends a set of policies designed to decrease the risk presented by AGI.

1 Introduction to AGI

> "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks."
>
> -Stephan Hawking

Superintelligent AI is a theoretical class of Artificial Intelligence that is "smarter than the best human brains in practically every field [1]." The successful development of Superintelligent AI would have a tremendous impact on humanity, perhaps more so than any invention in human history [2]. Intelligence has allowed humans to use tools and strategies which have enabled us to accomplish far more than any other animal. Our intelligence has also given us far greater destructive capabilities–we stand alone in the capability to destroy all major life on the planet. As we have unique abilities unfathomable to lesser life, a superintelligent AI would surpass us in a similar manner, accomplishing far more than is possible on our own. However, as human intelligence has given us the power to destroy, this superintelligent agent would be capable of far greater destruction than humans have comprehended.

Artificial General Intelligence (AGI) is the "intelligence of a machine that could perform any intellectual task that a human being can [3]." AI researchers believe that the development of AGI will quickly lead to Superintelligent AIs [2], meaning that the risk factors of a superintelligent AI apply fully to the development of AGI.The "Singularity" was originally described as the "ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue [4]." In other words, a hypothetical point in time when AI causes a runaway explosion of technological progress that will dramatically alter life on the planet.

Friendly Artificial Intelligence (FAI) is defined as "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 [5]." AGI systems that are also FAI systems are the ultimate goal of AI research, as an AGI system that is not "friendly" is dangerous, while an FAI system that is not "general" is limited.

Despite the existential risk that Superintelligent AI may pose [2], politicians have shied away from the issue [6], leaving the United States without a real policy. Likewise, there is a lack of awareness and a lack of concern among the general public [6]. While awareness of the risks within the AI research community has improved in recent years [7], there is still far more time, money, and effort spent on the research and development of the capabilities of AI than on its "friendliness."

1.1 Benefits of AGI

Nick Bostrom, philosophy professor at Oxford and Director the Future of Humanity Institute, claims that "superintelligence is radically different" from any other technology, and that "it would be much better at doing scientific research and technological development than any human, and possibly better even than all humans taken together [2]." The implications of this are staggering. One benefit of Artificial Intelligences is that they are easily replicated, meaning that once a superintelligent research AI is created, it could quickly be turned into a large arsenal of the greatest scientists the world has ever seen. Bostrom suggests that this could result in world-changing breakthroughs such as "the elimination of aging and disease [2]," with others suggesting that it would eliminate the demand for people to work [8].

1.2 Risks of AGI

It may be initially unclear how an AGI may come to harm humanity, when computer programs are known for doing merely what they are programmed to do. This fact is more dangerous than it initially seems. To demonstrate the danger in this, Oxford professor Nick Bostrom introduced a thought experiment involving a theoretical "Paperclip Maximizer [2]."

The Paperclip Maximizer is hypothetical AGI that is programmed with one goal: to maximize the number of paperclips in its inventory. It starts off innocently enough, learning how to find, buy, or trade for paperclips. At some point, it learns to manufacture paperclips on its own. But it is not bounded by morality; it may decide to steal or kill for paperclips, if it expects a positive outcome–any obstacle towards it achieving more paperclips is expendable. The real danger comes when it realizes the biggest obstacle to its own success: its own intelligence. Thus the Paperclip Maximizer begins a process of recursive self-improvement, leading to the theorized "intelligence explosion [2]." The principle is simple: once an AGI exceeds human intelligence, it is capable of programming itself better than the humans who created it did. Thus it becomes even more intelligent, and is in turn able to create an even smarter version of itself, and even create better hardware for itself, until the point where it is magnitudes smarter than any human. At this point, the Paperclip Maximizer is truly dangerous, as a superintelligence that will create paperclips at any cost.

Bostrom hypothesized that in time the entire world would be converted into paperclips, ending humanity and any other intelligent life permanently [2].

In practice it is unlikely that anyone would program an AGI for the purpose of maximizing paperclips, but it is argued that almost any goal would lead to similar actions: while not driven by a lust for power or knowledge in a way that a human might be, for almost any conceivable goal knowledge and power are a huge assets, thus any goal-driven AGI is likely to attempt recursive self-improvement to raise its intelligence, and gain any kind of "power," physical, social, political, by any means it can, without regard to morality [2]. Thus, the importance of programming Friendly AI becomes clear.

The Friendly AI issue is non-trivial, and researchers have argued that naive attempts at programming morality into machines will be ineffective [5]. A related example to the Paperclip Maximizer is the Smiley-Face Maximizer [9]. The example goes as follows: a researcher intent on making an AGI that will work for the generalized goal of the benefit of humanity attempts to teach the AGI what it means to make humans happy. To do this, he uses training data that includes pictures of humans smiling. The AGI seems to understand. Through a process similar to that used by the Paperclip Maximizer, the Earth is converted into disembodied smiley-faces.

Some have proposed a simple solution to the problem of AGI Risk: why not just contain the AGI until it is proven to be safe? FAI researcher Eliezer Yudkowsky has argued that this too will be ineffective. He first cites the fact that an AGI will want to leave the box to accomplish its goals, and therefore it will purposefully deceive its operators into thinking it is safe [10]. He further argues that a superintelligence will be intelligent enough to persuade its operators to let it out of the box. Thus, even a *contained* AGI poses a potential risk to humanity [10].

1.3 AGI Timeline

The expected time before AGI is devloped has a substantial impact on AGI policy. MIRI Executive Directer Luke Muehlhauser says that if it occurs within the coming decades, it is very important to begin working managing the risks immediately, while if is a few thousand years away, there are more important issues to be working on [11]. The natural approach for determining the timeline is to seek expert opinion.

Harvard professor and popular science author Steven Pinker claimed in 2008 that "There is not the slightest reason to believe in a coming singularity," comparing it to "jet-pack commuting" and "nuclear-powered automobiles" as unrealistic futures technologies that have been predicted for years, but never happened [12]. The sentiment seems to be shared by many other experts [12], yet Stuart Armstrong of the Future of Humanity Institute said that at the 2012 Singularity Summit the median predicted year of the Singularity was 2040 [13]. In the same year, Armstrong worked on a study published by MIRI about expert predictions which concluded that "AI timeline predictions have all the hallmarks of tasks on which [experts] would perform badly" and that there is "no indication that experts brought any added value when it comes to estimating AI timelines" over non-experts [14].

In other words, not only are the expert opinions highly varied, but for questions such as this one there experts do no better than the average person. MIRI

is continuing to perform research on possible methods for forecasting the Singularity, but for the moment we are somewhat ignorant of the likely timescale [11]. This makes the ideal course of action is somewhat unclear. However, in a 2013 MIRI paper, Yampolskiy and Sotala offered the following:

> It would be a mistake, however, to leap from "AGI is very hard to predict" to "AGI must be very far away." Our brains are known to think about uncertain, abstract ideas like AGI in "far mode," which also makes it feel like AGI must be temporally distant, but something being *uncertain* is not strong evidence that it is *far away*. When we are highly ignorant about something, we should widen our error bars in both directions. Thus, we shouldn't be highly confident that AGI will arrive this century, and we shouldn't be highly confident that it *won't* [15].

This suggests that while a timeline for the development of AGI is uncertain, given the extreme risks presented by AGI, research on AGI risks and FAI is important *now*, due the possibility of AGI being developed in the near-future.

2 Organizations Involved in AGI Risk

The majority of research of AGI risk and Friendly AI is done by a small number of non-profit organizations. Some of these organizations also promote the awareness of AGI risk among the AI research community and the general public. This section details the organizations most heavily involved in the field.

2.1 Machine Intelligence Research Institute

The Machine Intelligence Research Institute (MIRI) is a non-profit research group devoted to "Ensuring that the creation of smarter-than-human intelligence has a positive impact [16]." They are one of the only groups actively exploring the technical questions surrounding "trustworthy" designs for superintelligent AIs. They claim to be the only group offering full-time positions in Friendly AI research [17]. The core of their stance is that the current technological climate which "favors the incremental development of algorithms that are not particularly transparent, robust, or stable" is not suitable for the development of safe AIs, and that research should be done now in order to lay the groundwork for Friendly AI [16]. According to MIRI's technical agenda [40]:

> It is prudent to develop a theory of superintelligence alignment before developing a system capable of attaining or creating superintelligence. It may seem premature to tackle the problem now, with superintelligent systems still firmly in the domain of futurism. But imagine the chagrin if, in a few decades, the need for a mature theory of corrigibility is imminent, but the field is just as immature as seen in this technical agenda!
>
> We think it is wise to approach these problems as soon as they look approachable. To do otherwise seems to us like a cognitive bias surrounding the fear of wasted effort, rather than a prudent

calculation of the probable consequences of doing something versus nothing.

Despite being highly invested and highly concerned with superintelligent AI, MIRI has somewhat distanced itself from policy concerns, instead choosing to focus on a research agenda. In a 2013 MIRI paper, research associate Kaj Sotala said that they were "generally supportive of regulation, though the most effective regulatory approach remains unclear [15]." MIRI is funded mostly through private donations and fundraising, but claims to be making "efforts to find and apply for grants from both private and public grantmakers [17]."

2.2 Future of Humanity Institute

The Future of Humanity Institute (FHI) is a research institute at Oxford which "enables a select set of leading intellects to bring the tools of mathematics, philosophy, and science to bear on big-picture questions about humanity and its prospects [18]." FHI's focus is primarily on "existential risks" to humanity, which FHI Director Nick Bostrom defines as "[a risk] where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential [19]." AGI is among these risks, as well as topics such as bioterrorism, human enhancement, and nanotechnology.

2.3 Centre for the Study of Existential Risks

The Centre for the Study of Existential Risks (CSER) is a Cambridge-based organization which, like the FHI, performs research on various existential risks that may face humanity, but with a particular focus on AI [20]. It was founded in 2012 by Skype co-founder Jaan Tallinn, Bertrand Russell Professor of Philosophy Huw Price, and University of Cambridge Emeritus Professor of Cosmology & Astrophysics Martin Rees [20]. According to CSER's website, their "goal is to steer a small fraction of Cambridge's great intellectual resources, and of the reputation built on its past and present scientific pre-eminence, to the task of ensuring that our own species has a long-term future [21]."

2.4 Future of Life Institute

The Future of Life Institute (FLI), founded in 2014, is a new-comer to the field of AGI risk [22]. It was founded by MIT cosmologist Max Tegmark and Skype co-founder Jaan Tallinn and others [22], with the intent to "mitigate existential risks facing humanity" with a special focus on Artificial Intelligence [23]. The FLI made waves when billionaire Elon Musk made a donation of 10 million dollars [24]. Thanks to the generous donation, and a high-profile team of talent, they appear poised to make a substantial impact. It has very close ties to CSER, as they are both located and Cambridge and there is substantial overlap between the personnel [25]. When asked about what differentiates the the FLI from MIRI, CSER, and FHI, FLI co-founder Viktoriya Krakovna said "Compared to FHI and CSER, we are less focused on research and more on outreach, which we are well-placed to do given our strong volunteer base and academic connections. Our location allows us to directly engage Harvard and MIT researchers in our brainstorming and decision-making [26]."

2.5 Association for the Advancement of Artificial Intelligence

The Association for the Advancement of Artificial Intelligence (AAAI) is, according to their website, "a nonprofit scientific society devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines [27]." The AAAI was founded in 1979 to promote research and awareness of various AI topics [28]. They publish the quarterly magazine *AI Magazine* [29], and organize the "AAAI Conference on Artificial Intelligence," which is considered one of the top conferences in the field of AI [30].

In 2009, the AAAI held the "Asilomar Meeting on Long-Term AI Futures," which studied the potential future risks of AI [31], and raised awareness of these issues among AI researchers [7]. According to the AAAI, the location of Asilomar was chosen because it "resonated broadly with the 1975 Asilomar meeting by molecular biologists on recombinant DNA—in terms of the high-level goal of social responsibility for scientists [31]."

3 Current Actions

There as of yet has been no regulatory legislative action to protect against AGI risks [15], and the topic has received little or no interest politically [6]. Most of the actions taken thus far have been either the promotion of awareness of the risks, or direct research into the risks and possible solutions.

3.1 Awareness

The first issue of managing the risks of AGI is raising awareness of these risks. Awareness of these risks has been improving in the AI research community, in a part due to the 2009 AAAI Asilomar meeting [7]. Due to an increase in donations [32] and improved management following the hiring of Luke Muelhauser [33], the research output of MIRI drastically increased in the following years [34], there does not seem to have been a substantial increase in the research output outside of MIRI. MIRI has made continuing efforts to increase awareness, most notably by running workshops focus on various issues surrounding open problems in Friendly AI [35]. To create and run these workshops, they have enlisted the help of individuals from top universities as well as top tech companies [35].

Unfortunately, the efforts of the AAAI and MIRI have done little to improve awareness *outside* of the research community [7]. Several Famous individuals who made their livings in Science and Technology fields have gone public with their concern for AGI risks, including Bill Gates, Stephan Hawking, and Elon Musk [6], but despite this Dave Atkins says that "Worrying about artificial intelligence, then, is inconvenient for both sides of the political aisle, and therefore gets waived off as the province of nerds too nerdy even for politics [6]." Thus, the overwhelming majority of both politicians and the general public are too unmotivated and uninterested in the issue to aid in any progress.

MIRI, (the Singularity Institute at the time), has taken an extremely unorthodox approach towards solving the problem of public indifference. Seeing

the public's reaction's indifference towards the potentially enormous AGI risks as a failure of rationality, a failure of the human thinking process itself, they created in 2012 the Center for Applied Rationality (CFAR) [36]. CFAR has already begun making in impact: they've already conducted corporate workshops for big-name companies such as Facebook [36]. The biggest name in the movement is MIRI Senior Research Fellow Eliezer Yudkowsky, creator of LessWrong, which is a website self-described as "a large, active website for people who try to think rationally [37]." In an extremely bizarre but ultimately successful move, to promote Rationality Yudkowsky wrote what became the most popular piece of FanFiction in the world: *Harry Potter and the Methods of Rationality* [38].

3.2 Research

Thinking about the risks of AGI and ways to avert them dates back a long ways–Asimov's famous "Three Laws of Robotics" were original published in a 1942 short story titled "Runaround [39],"which while now widely are considered to be "not a viable approach for safe AI [15]," where important in bringing the concept of AGI risk to awareness. However, research in earnest did not begin until MIRI's first publication in 2001, "Creating Friendly AI 1.0," which coined the term "Friendly AI" and introduced many of the key concepts in the field [5]. The perhaps most noteworthy outcome of the paper was the realization that creating Friendly AI would be *difficult*, in a way separate from the difficulty of creating AGI itself, and the corresponding realization that FAI research needed to be done before AGI was imminent in order to avoid the associated risks [5].

The Future of Humanity Institute joined the fold in 2005 [18]. The first related paper was written by Director Nick Bostrom, titled "Existential Risks," and was published in 2002 [19]. The paper coined the term "existential risk," which was previously defined in this paper. While early in its history, FHI did not focus as heavily on issues specifically surrounding AGI, much of their recent research has focused on the topic [42]. In 2014, the Oxford University Press published Bostrom's book, *Superintelligence: Paths, Dangers, Strategies* [1], which FHI calls "the most comprehensive work to date detailing the existential risks of artificial superintelligence [42]."

4 Possible Approaches

The section examines broadly four possible courses of action that may be taken towards managing AGI risk, and the possible results of those courses of action. The first is the "Do Nothing" approach, in which no research into AGI risks or FAI is performed, and no regulation is in place. The second approach is "Independent Research," in other words the course we are on right now, in which independent organizations not receiving public funds continue to undertake research and promotion of awareness. The third course is the "Public Funding" approach, in which the government promotes and funds research of AGI risks and safe AI systems. The fourth course is "Strict Regulation," where the government imposes strict regulations on and the research and development of AI systems.

4.1 Do Nothing

Many proponents of this approach argue that there is in fact nothing to be concerned about, and therefore nothing needs to be done. Alan Winfield, an Electric Engineering professor at the University of the West of England, in an article for the *The Guardian* claimed that "Artificial Intelligence will not turn into a Frankenstein's monster [41]." He compared it to Faster-Than-Light travel in its difficulty to achieve, which in turn makes the risk negligible. He further claimed that an existential threat from AI is even *more* unlikely because the number of things he perceives need to go right in order for it to occur:

> "For the risk to become real, a sequence of things all need to happen, a sequence of big ifs. If we succeed in building human equivalent AI and if that AI acquires a full understanding of how it works, and if it then succeeds in improving itself to produce super-intelligent AI, and if that super-AI, accidentally or maliciously, starts to consume resources, and if we fail to pull the plug, then, yes, we may well have a problem. The risk, while not impossible, is improbable [41]."

Perhaps the greatest proponent of the "Do Nothing" strategy of response is the societal inertia that has opposed new discoveries and topics in science for centuries, holding a seemingly higher burden of proof than even the strictest of scientific community. Clear evidence can be seen in the history of the evolution debate–Charles Darwin's famous *On the Origin of the Species* was published 1859, and *to this day* there are states which *require* that students "critically analyze key aspects of evolutionary theory [43]." This is in a field where there is *no debate* among scientists.

However, MIRI Senior Research Fellow Eliezer Yudkowsky says that it is even worse than that, claiming the "general sanity waterline is currently *really ridiculously low*. Even in the highest halls of science [44]," suggesting that even with strong proof of the risks of AGI, there still may be little or no reaction from society or the scientific community at large, and even if the scientific community reaches a consensus, there is reason to believe that politics will block any significant action. For evidence of this, consider the subject of Climate Change: by 2001 there was already a consensus among scientists that human emissions were the cause of the global temperature increase [45], yet the United States failed to ratify the Kyoto Protocol, with President George W. Bush citing concerns that it would hurt the US economy [46]. The economic impact of AGI has the potential to be orders of magnitude greater than any impact the Kyoto Protocol could have possibly had. So as long as there are those who prefer to risk destruction of the world over marginal economic impacts, there will be opponents of imposing any sort of restrictions on AGI.

4.2 Independent Research

Currently FAI research is carried out by what Gary Marcus of *The New Yorker* calls "A tiny cadre of brave-hearted souls," noting that "annual amount of money being spent on developing machine morality is tiny [47]." It is essentially a charity act, carried out by the few who have realized the risk: the Future of Life Institute for instance is funded primarily by a 10 million dollar donation made by Elon Musk [24]. Without any government policies, this is likely to be

how the research is sustained. There is little financial benefit to be had through FAI research, meaning without grant money researchers must rely on donations by concerned individuals.

While it is possible that the research performed by these organizations before the creation of AGI will be sufficient, it seems questionable to rely on the generosity of Silicon Valley billionaires to protect the future of humanity. It is also a very passive way of addressing the issue, and runs the risk of becoming a matter of "too little, too late."

4.3 Public Funding

John McGinnis, professor at Northwestern University School of Law, in a paper advocating for the government to assist in the "acceleration" of the development of AI, suggested implementing a system similar to the National Institutes of Health, offering incentives and grants based on a peer review process [48]. He recommends that "Peer review panels of computer and cognitive scientists would sift through projects and choose those that are designed both to advance AI and assure that such advances would be accompanied by appropriate safeguards [48]." He also notes such an institution would be "quite modest and inexpensive" at first, and that it could be expanded once it proved itself [48].

Given the tiny amount of funding that FAI research currently receives, setting up such an institution certainly seem like it would be beneficial to the cause. Offering incentives to AI researchers who specifically pursue "safe" approaches increases the likelihood that AGI will be developed safely. The NIH also has to frequently deal with issues that could affect public safety, including some that could arguably fall under the category of Existential Risk, or at least Global Catastrophic Risk. Therefore, drawing a parallel to the NIH and the field of medicine makes sense on a basic level.

However, there are issues unique to the field of AI that may make simple control of funding ineffective at managing AGI risks, compared to the managing the risks of biomedical research. Perhaps the most significant is the fact that AI research is not nearly as dependent on funding as biomedical research is. It is estimated that the average cost of developing a new prescription drug is now greater than $2.5 billion [49]. Compare this to the cost of the development of IBM Watson, a cutting-edge AI computer system: according to the technical article, it took twenty engineers three years to build, with the hardware costing approximately $3 million dollars [50]. While the engineer salaries and infrastructure costs are unknown, the initial price-tag was on the order of tens of millions, not billions. For big companies and research institutions, this cost is affordable, with or without grants or funding, meaning a company interested in developing an unsafe AGI without approval from the review panel would be able to do so.

Another considerable issue is that creating an NIH-like institution in the US would be limited in scope to the US. It is true that modern markets are global, and that promoting safe AI research in the United States would have an impact on the global market, but it is possible that this alone will not be enough.

For these reasons, it's unclear whether or not AGI risks could be effectively managed by an NIH-like institution. It is possible that a higher level of regulation is necessary.

4.4 Strict Regulation

The highest profile global risk in the 20th century was likely the development of Nuclear Weapons. Their potential was realized quickly, and it wasn't long before campaigns began to limit their spread and development. Many consider a future free of nuclear weapons to be the safest future for humanity. For that reason, regulation on nuclear weapons transcends borders. The Treaty on the Non-Proliferation of Nuclear Weapons is an international treaty joined by 190 countries, which seeks the non-proliferation and disarmament of nuclear arsenals [51]. In a 1957 treaty, the International Atomic Energy Agency was created "to promote the safe, secure and peaceful use of nuclear technologies [52]."

Perhaps the risks of AGI are similar to the risks of nuclear energy: safe use of both can lead to substantial rewards for society, but they both have the potential to be catastrophic to humanity. Elon Musk has claimed that AGI is an even bigger risk than nuclear weapons [53]. However, as difficult as managing the spread and creation of nuclear weapons has been, preventing the spread of AGI will be even harder.

The reasons for this are simple: all it takes to develop an AGI system is computing power, engineers, and researchers. While the process for creating such a system is not yet known, as soon as it is known, anyone with the knowledge will be able to create such a system. While the basics of a nuclear bomb are well known, creating functional weapons still involves access to weapons-grade uranium, extensive knowledge of the refining process, and the development of an effective delivery system [54], especially if you intend to get around modern missile and air defense systems. These obstacles likely do not exist for AGI: the only resources needed will be the design, computing hardware, and the engineers. One thing no government is likely to do is limit the spread of computing power.

So perhaps the last resort is to enforce strict regulations on AGI research, disallowing any research that is not considered safe. Sotala and Yampolskiy in 2013 produced a survey of possible responses to AGI risks, which included a number of possible technical solutions that could be mandated by government regulations [15]. The problem with this is once again: enforcing technical regulations would be difficult. Organizations willing to bend the rules would be able to easily carry out research in secret, or even release products that do not implement certain safety protocols, as it may be difficult to detect non-compliance without examining the source cod.e The potential rewards and prestige of developing an AGI system are immense, without Draconian measures being put into place it seems unlikely that none would attempt it, and once the first AGI systems are developed, it may be to the perceived benefit of companies and organizations to ignore regulations to improve the efficiency of their products.

The comparison to nuclear technologies clarifies another potential issue with trying to regulate AGI: despite widespread awareness of the dangers of nuclear technologies, and international treaties designed to ensure the nuclear technology is used for the good of humanity and not the harm, there have still been a number of disasters related to nuclear energy. The initial use of nuclear power was offensive, as part of the Manhatten Project. If AGI debuts as an agent of war, the risks are innumerable. However even positive uses of nuclear power have led to disaster, the highest profile incident being the Chernobyl disaster [55]. The Chernobyl disaster was caused by mistakes from the operators and

engineers at the power plant[55]. Mistakes involving AGI could cause disasters at a much greater scale.

5 Policy Recommendations

After investigating the topic of AGI risk, and examining various suggestions and proposals, I recommend a three-pronged policy approach towards managing AGI risk: 1) Establish public funding of AGI risk and FAI research, 2) Establish an international agency devoted to the safe development of AGI, and 3) Encourage continuing independent efforts to research and promote the awareness of AGI risk and FAI.

5.1 Federal Friendly AI Research Funding

I believe that the first step from a policy perspective towards mitigating the risks of AGI is to begin funding research on federal dollars. I agree with McGinnis's proposal for setting up a small peer-review panel which funds research into Friendly AI and other topics related to AGI risk management [48]. The initial cost would be small, and the potential benefits high. MIRI has been able to perform valuable work on a budget of well under $2 million dollars per year [32], but as they are performing work that is beneficial to the public, it is unfair that they are funded by a small number of donors.

This would also give the benefit of allowing new minds a chance to compete for grant money, diversifying the pool of researchers. Currently almost all FAI research is performed by a very small number of organizations who subsist mostly off of donations. It is time to open the field of FAI to the larger research community.

5.2 International Agency

As the risks of AGI affect the entire world, and not just the United States, I also am supportive of efforts to establish an international treaty comparable to the IAEA Statute. The IAEA promotes the peaceful use of nuclear energy, and like nuclear energy, AGI can contribute positively to humanity. If done properly, AGI will be mankind's greatest invention. Therefore, it is important that the international community comes together to ensure that AGI is a benefit to humanity.

5.3 Continued Independent Research and Awareness Efforts

While perhaps not an issue of public policy, I believe that it is important for organizations to continue their efforts to promote the awareness of AGI risks. There will not be support for AGI policies at the Federal or International level until a larger segment of the population begins to take AGI risks seriously. It is currently treated by many as a fringe issue that is not to be taken seriously, or as an issue that is too far in the future to be concerned with. This is in spite of research suggesting that FAI is a serious issue that needs to be addressed

properly in advance of the development of AGI [5], as well as research suggesting that we will not be able to accurately forecast the arrival of AGI [56].

Because of the difficulties of implementing and enforcing strict regulations on AGI, the issue of AGI risk awareness among those creating AGIs is even more important. I do not believe any company will consider it inside of their interests to create an intelligence that threatens humanity. If they are given the tools to create safe AGIs, through FAI research and awareness, companies should be able to use AI to create products that are beneficial to humanity.

6 Summary

The enormity of the risks posed by the development of superintelligent AGI are balanced only by the potential rewards. MIRI, the FHI, the AAAI, CSER, the FLI, and others are continuing to work on both research into how to mitigate the risks of AGI, and spreading awareness among both researchers in the field of AI and the general public. There are a number of potential pitfalls to be weary of regarding public policy on AGI, but public funding of FAI research and an international push towards the development of safe AI will improve the chances of success in navigating this risk, and the chances of reaping the rewards of what will be humanities greatest invention.

References

[1] Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* New York: Oxford University Press.

[2] Bostrom, Nick. "Ethical Issues in Advanced Artificial Intelligence." Nick Bostrom. 2003. Web. 6 Mar. 2015.http://www.nickbostrom.com/ethics/ai.html

[3] Ryan, Michael. "Digital Mind: An Exploration of artificial intelligence." CreateSpace Independent Publishing Platform. 2014. 9.

[4] Ulam, Stanislaw (May 1958). "Tribute to John von Neumann". 64, #3, part 2. Bulletin of the American Mathematical Society. p. 5. https://docs.google.com/file/d/0B-5-JeCa2Z7hbWcxTGsyU09HSTg/edit?pli=1

[5] Yudkowsky, Eliezer. "Creating Friendly AI-The Analysis and Design of Benevolent Goal Architectures." (2001). https://intelligence.org/files/CFAI.pdf

[6] Atkins, David. "If Bill Gates, Elon Musk and Stephen Hawking Are Worried, Shouldn't You Be?" The Washington Monthly. 8 Feb. 2015. Web. 6 Mar. 2015. http://www.washingtonmonthly.com/political-animal-a/2015_02/if_bill_gates_elon_musk_and_st054073.php

[7] Muehlhauser, Luke. "Diana Spears on the Safety of Adaptive Agents." Machine Intelligence Research Institute. 9 Apr. 2014. Web. 6 Mar. 2015. https://intelligence.org/2014/04/09/diana-spears/

[8] Hanson, Robin. "IF UPLOADS COME FIRST." 8 Mar. 1994. Web. 6 Mar. 2015. http://mason.gmu.edu/~rhanson/uploads.html

[9] Yudkowsky, Eliezer. "Artificial intelligence as a positive and negative factor in global risk." Global catastrophic risks 1 (2008): 303. http://intelligence.org/files/AIPosNegFactor.pdf

[10] Yudkowsky, Eliezer. "The AI-Box Experiment:." Yudkowsky. 2002. Web. 6 Mar. 2015. http://www.yudkowsky.net/singularity/aibox

[11] Muehlhauser, Luke. "When Will AI Be Created?" Machine Intelligence Research Institute. https://intelligence.org/2013/05/15/when-will-ai-be-created/

[12] "Tech Luminaries Address Singularity." IEEE Spectrum. 1 June 2008. Web. 6 Mar. 2015. http://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity

[13] Armstrong, Stuart. "Stuart Armstrong: How We're Predicting AI." FORAtv Video. Web. 6 Mar. 2015. http://library.fora.tv/2012/10/14/Stuart_Armstrong_How_Were_Predicting_AI

[14] Armstrong, Stuart, and Kaj Sotala. "How we're predicting AI–or failing to." Beyond Artificial Intelligence. Springer International Publishing, 2015. 11-29. https://intelligence.org/files/PredictingAI.pdf

[15] Sotala, Kaj, and Roman V. Yampolskiy. "Responses to catastrophic AGI risk: a survey." Physica Scripta 90.1 (2015): 018001. http://intelligence.org/files/ResponsesAGIRisk.pdf

[16] "Overview - Machine Intelligence Research Institute." https://intelligence.org/files/MIRI_Overview.pdf

[17] Muehlhauser, Luke. "Our Mid-2014 Strategic Plan." Machine Intelligence Research Institute. 11 June 2014. Web. 6 Mar. 2015. https://intelligence.org/2014/06/11/mid-2014-strategic-plan/

[18] "About — Future of Humanity Institute — Programmes." Future of Humanity Institute. Web. 6 Mar. 2015. http://www.oxfordmartin.ox.ac.uk/research/programmes/future-humanity/

[19] Bostrom, Nick. "Existential Risks." Journal of Evolution and Technology. 9 Mar. 2002. Web. 6 Mar. 2015. http://www.jetpress.org/volume9/risks.html

[20] Lewsey, Fred. "Humanity's Last Invention and Our Uncertain Future." University of Cambridge. Web. 6 Mar. 2015. http://www.cam.ac.uk/research/news/humanitys-last-invention-and-our-uncertain-future

[21] "About." Centre for the Study of Existential Risk. Web. 6 Mar. 2015. http://cser.org/about/

[22] Chen, Angela. "Is Artificial Intelligence a Threat?" The Chronicle of Higher Education. 10 Sept. 2011. Web. 6 Mar. 2015. http://chronicle.com/article/Is-Artificial-Intelligence-a/148763/

[23] Moyer, Justin. "Elon Musk, Stephen Hawking, Google Researchers Join Forces to Avoid 'pitfalls' of Artificial Intelligence." Washington Post. The Washington Post, 12 Jan. 2015. Web. 6 Mar. 2015. http://www.washingtonpost.com/news/morning-mix/wp/2015/01/12/elon-musk-stephen-hawking-google-execs-join-forces-to-avoid-unspecified-pitfalls-of-a

[24] "Elon Musk Donates $10M to Keep AI Beneficial." Future of Life Institute. 15 Jan. 2015. Web. 6 Mar. 2015. http://futureoflife.org/misc/AI

[25] "Who We Are." Future of Life Institute. Web. 6 Mar. 2015. http://futureoflife.org/who

[26] Krakovna, Viktoriya. "Vika comments on New organization." LessWrong. http://lesswrong.com/lw/kcm/new_organization_future_of_life_institute_fli/b06p

[27] Association for the Advancement of Artificial Intelligence. Web. 6 Mar. 2015. http://www.aaai.org/home.html

[28] Waltz, David. "An Opinionated History of AAAI." AI Magazine 26.4 (2005): 45. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1845/1743

[29] "About the Journal." AI Magazine. Web. 6 Mar. 2015. http://www.aaai.org/ojs/index.php/aimagazine/about

[30] "Top Conferences in Artificial Intelligence." Microsoft Academic Search. Web. 6 Mar. 2015. http://libra.msra.cn/RankList?entitytype=3&topDomainID=2&subDomainID=5&last=0&start=1&end=100

[31] "Interim Report from the Panel Chairs." Association for the Advancement of Artificial Intelligence. Web. 6 Mar. 2015. http://research.microsoft.com/en-us/um/people/horvitz/note_from_AAAI_panel_chairs.pdf

[32] "Transparency & Financials." Machine Intelligence Research Institute. Web. 6 Mar. 2015. https://intelligence.org/transparency/

[33] Helm, Louie. "Interview with New MIRI Research Fellow Luke Muehlhauser." Machine Intelligence Research Institute. 14 Sept. 2011. Web. 6 Mar. 2015. https://intelligence.org/2011/09/15/interview-with-new-singularity-institute-research-fellow-luke-muehlhuaser-september-2

[34] "All Publications." Machine Intelligence Research Institute. Web. 6 Mar. 2015. https://intelligence.org/all-publications/

[35] "Research Workshops." Machine Intelligence Research Institute. Web. 6 Mar. 2015. https://intelligence.org/workshops/

[36] Segran, Elizabeth. "Inside The Rationality Movement That Has Silicon Valley Buzzing With Positive Thinking." Fast Company. 21 Oct. 2014. Web. 6 Mar. 2015. http://www.fastcompany.com/3037333/most-creative-people/inside-the-rationality-movement-that-has-silicon-valley-buzzing-with-po

[37] "About LessWrong." LessWrong. Web. 6 Mar. 2015. http://lesswrong.com/about/

[38] Whelan, David. "The Harry Potter Fan Fiction Author Who Wants to Make Everyone a Little More Rational." VICE. 2 Mar. 2015. Web. 6 Mar. 2015. http://www.vice.com/read/theres-something-weird-happening-in-the-world-of-harry-potter-168

[39] Asimov, Isaac. I, robot. Spectra, 2004.

[40] Soares, Nate, and Benja Fallenstein. *Aligning Superintelligence with Human Interests: A Technical Research Agenda.* Tech. rep. Machine Intelligence Research Institute, 2014. URl: http://intelligence.org/files/TechnicalAgenda. pdf, 2014.

[41] Winfield, Alan. "Artificial Intelligence Will Not Turn into a Frankenstein's Monster." The Guardian. 9 Aug. 2014. Web. 6 Mar. 2015. http://www.theguardian.com/technology/2014/aug/10/artificial-intelligence-will-not-become-a-frankensteins-monster-ian-winfield

[42] "Publications." Future of Humanity Institute. Web. 6 Mar. 2015. http://www.fhi.ox.ac.uk/research/publications/

[43] "Texas Improves on Strengths and Weaknesses Language in Science Standards on Teaching Evolution." Discovery Institute. 27 Mar. 2009. Web. 6 Mar. 2015. http://www.discovery.org/a/9851

[44] Yudkowsky, Eliezer. "Raising the Sanity Waterline." Less Wrong. 12 Mar. 2009. Web. 6 Mar. 2015. http://lesswrong.com/lw/1e/raising_the_sanity_waterline/

[45] McCarthy, James J., ed. Climate change 2001: impacts, adaptation, and vulnerability: contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 2001. http://www.grida.no/publications/other/ipcc_tar/

[46] Dessai, Suraje. "The climate regime from The Hague to Marrakech: Saving or sinking the Kyoto Protocol." Tyndall Centre for Climate Change Research Working Paper. TC f. C. C. Research. Norwich 27 (2001). www.tyndall.ac.uk/sites/default/files/wp12.pdf

[47] Marcus, Gary. "Moral Machines." The New Yorker. 24 Nov. 2012. Web. 6 Mar. 2015. http://www.newyorker.com/news/news-desk/moral-machines

[48] McGinnis, John O. "Accelerating AI." Nw. UL Rev. 104 (2010): 1253. http://www.law.northwestern.edu/LAWREVIEW/Colloquy/2010/12/

[49] Mullin, Rick. "Cost to Develop New Pharmaceutical Drug Now Exceeds $2.5B." Scientific American Global. 14 Nov. 2014. Web. 6 Mar. 2015. http://www.scientificamerican.com/article/cost-to-develop-new-pharmaceutical-drug-now-exceeds-2-5b/

[50] Anderson, Michael, et al. "Brainwash: A Data System for Feature Engineering." CIDR. 2013. http://www.cs.stanford.edu/people/chrismre/papers/mythical_man.pdf

[51] "Non-Proliferation of Nuclear Weapons (NPT)." UN News Center. Web. 6 Mar. 2015. http://www.un.org/disarmament/WMD/Nuclear/NPT.shtml

[52] "History of the IAEA." IAEA. Web. 6 Mar. 2015. https://www.iaea.org/about/history

[53] Newcomb, Alyssa. "What Elon Musk Says Could Be More Dangerous Than Nuclear Weapons." ABC News. ABC News Network, 4 Aug. 2014. Web. 6 Mar. 2015. http://abcnews.go.com/Technology/elon-musk-dangerous-nuclear-weapons/story?id=24830269

[54] Nyugen, Tuan. "Why It's So Hard to Make Nuclear Weapons." LiveScience. TechMedia Network, 22 Sept. 2009. Web. 6 Mar. 2015. http://www.livescience.com/5752-hard-nuclear-weapons.html

[55] "Chernobyl Accident 1986." World Nuclear Association. Web. 6 Mar. 2015. http://www.world-nuclear.org/info/Safety-and-Security/Safety-of-Plants/Chernobyl-Accident/

[56] Armstrong, Stuart, and Kaj Sotala. "How we're predicting AI–or failing to." Beyond Artificial Intelligence. Springer International Publishing, 2015. 11-29. https://intelligence.org/files/PredictingAI.pdf