Final Project

College of Science and Engineering
University of Minnesota

Name:	Robert Niemann
Student Number:	5457661
Course:	CSci 5512
Research Project Title:	Setting the Stage for Sustainable AI
Professor:	James Parker

1 Abstract

For this paper, I introduce some of the hot topic issues of AI (such as the advent of super-intelligence, the treatment of collected data, the AI arms race), including information gathered from polling different sources as to what experts believe and have reported on what their opinions are on which future guidelines should be followed for AI, what guidelines are currently being followed and to what extent, as well as what current issues there exist in AI and what is being done about them. After the literature review, I will present my own thoughts (some of which listed in the introduction), and compare/contrast what the biggest issues in AI are, should be, and what should be done about them moving forward from my own beliefs to the information presented from different sources in this article.

2 Introduction

One of the fastest growing fields in technology is Artificial Intelligence, and for good reason. It has demonstrated the ability to improve many aspects of human life: self-driving cars, disease identification and user suggestions in online media to name a few, as well as the promise of solving problems that humans have not been able to thus far, such as curing cancer or discovering inter-planetary travel.

But among these promises lie similarly great risks, aligning with bright visions almost as if two sides of the same coin. A few of the biggest concerns of the advent of AI are listed below:

The singularity is the point in time in which the growth of AI is no longer controllable, or when humanity no longer makes the final decision in how machines behave. For this to happen, AI will necessarily become smarter than all of humanity, which whether an ultimately good or bad event, will most likely occur. Take the following logic for example: so long as humanity is able to, they will continue to create more intelligent computers as those intelligent computers will come with more power for the wielder. Therefore, either a catastrophic event will occur which causes humans to totally cease technological progress, or humans will develop machine intelligence as much as possible. Eventually, there will come a point where humans are not able to create more intelligent robots on our own, so we will need to develop machines to create even more intelligent machines. At this point, we will have lost control over how super-intelligent robots are developed, and have fulfilled the singularity prophecy.

Another potential disaster resulting from the hyper-advancement of AI stems from the fear of how much military-grade power an entity would gain from having a human-level AI. Because computers are so much faster than the human brain (around 10 million times so), if a country were to develop a human-level AI and run it for 6 weeks, giving it our current knowledge with the assumption that no other country has this capability, that country would have effectively advanced 500,000 years into the future, thoroughly outclassing any of its competitors. Not only would this spell bad news for any other countries in the world, but should those countries suspect the threat of such an intelligence they would have great incentive to destroy it, including but not limited to nuclear weaponry.

There are two large risks to the current American economic system that automation poses, affecting the 'higher' and 'lower' ends respectively. The first concern would arise from using automated 'brokers' (not necessarily of human-level intelligence) in order to trade stocks nearly instantaneously according to market price. While initially sounding like a good way to speed up the market, this could result in the market spinning out of control. The basis for this fear comes from the problem-solving ability we've seen from AI thus far: sometimes the most optimal way to solve an argument is not practical in the real-world, yet an AI might take that route anyway due to unforeseen oversights in whatever cost function it might be using. On the level of the stock-market, it is highly potential that two agents might find a way to manipulate the stock market to create some 'infinite money' glitch, which could result in the complete transfer of value from some company to some individual. This doesn't even account for people maliciously using AI to purposely mess with the market, as humans have found ways to abuse it in the past, there's no telling the number of breaks a human-level AI could find in our current system. All in all, high-level AI could pose a fantastic risk to the current stock market.

The second way in which AI could threaten the economy is in automating simpler, lower-class jobs. Automation has already taken some jobs, such as in Amazon's smart warehouses, and looks poised to take even more jobs, take for instance long distance trucking wherein professional are already more susceptible to accidents than self-driving cars. And who's to say that as AI develops, higher-class jobs won't also be taken? If A.I. can reliably replicate the intelligence found in humans, it would take a minute fraction of the time to train machines to do nearly any job that humans have spent years mastering. What laws are there to prevent the top 10 people in the world to completely and totally automate every other job there is and essentially own the economy, simply by having the right amount of money at the right time?

Another at-risk aspect of life would be that of privacy and security. First, AI has already made great strides in the realm of 'Deep-faking', in which a video is generated to make it look as though an individual were talking about whatever was specified by the creator. In a legal system in which video evidence is one of if not the primary basis for conviction (one major recent example being the People v. Rittenhouse case), when is evidence going to start being generated for crimes that were never committed? Separately, what does the advent of AI spell for our notion of digital security? Forget using quantum computing to totally dismantle the fundamentals of cyber-security; if humans have already been able to poke holes and break in to existing systems, how much more could a robot exploit those holes? Having someone's entire digital footprint at your disposal, as well as being to generate evidence in today's system would theoretically allow you to send whoever you wanted to prison so long as you made a convincing enough case with enough truth woven in, not to mention any bit of your personal data that you thought was private. Goodbye text message to Mom, hello hacker!

As you can see, there are numerous risks that could potentially come with the advent of super-intelligence. Ironically, these risks encompass the end goal for many organizations (Governments want ultimate military, criminals would love to hack anyone, the 1 percent would never turn down completely automated labor), so there is no slowing down the research and development of these computers. Therefore, the action taken to avoid these risks must come in the form of regulation and precaution, as once someone achieves this ultimate intelligence, they will more likely than not be calling the shots. In the next section, I survey the current state of affairs when it comes to what precautions have been taken thus far.

3 Related Work

In dealing with the possibility of super intelligence, it is important to understand the viewpoints of world leaders and how those would change once a Super-AI were introduced to the world. Instead of attempting to find a golden standard for how countries deal with AI, or explore multiple different standards and what they entail, a 'Public Policy Vector Field'[1] would map given importance on certain issues from before to after the realization of Super-AI, and specifically the direction countries are likely to want to follow as a guideline for implementing more concrete policies.

In the study, Bostrom indicated 4 major properties in which to measure reactions (properties being like dimensions on the vector field): efficiency, allocation, population and process. The four main parameters of efficiency that hold especially to incredible AI improvement would be technological opportunity, AI risk, global miscommunication probability and 'reducing turbulence'. In terms of efficiency, technological opportunity refers to the ability to fully capture the extreme growth in AI, and to make sure that the growth makes an impact in a timely manner such that much of the population present at its creation can reap the benefits of such AI. For example, extending human life may be possible with this intelligence, so efficient technological growth would mean realizing this in time for a good portion of the population to have the opportunity to have their lives extended. AI risk is fairly explanatory via some of the risks from the introduction, efficient implementation would entail avoiding such risks. Global miscommunication would occur in line with the 'arms race' idea specifically, having countries avoid developing world-ending weapons through the use of AI, and reducing turbulence entails making sure that our current systems are able to evolve with AI such that new systems don't pop up in their place creating inefficiency.

Allocation refers namely to the development of AI, including areas of risk externalities, 'reshuffling', the 'veil of ignorance', and cornucopia. The development of Super-AI would entail risks to people across the world, and as such the notion of risk externalities would make sure that those exposed to risk be compensated regardless of whether that risk comes to realization. Reshuffling refers to the potential shift in wealth which would occur due to a potential intelligence revolution, in which case the goal would be to have the wealth distribution hopefully stay as it is, if not becoming less concentrated. The veil of ignorance refers to the fact that nobody knows when or where this power will spawn, in which case the rights and freedom of every individual would hopefully be affected to a minimal extent upon one agent achieving power, if not becoming more liberated. Finally, cornucopia refers to the plan for whatever country that realizes immense benefits (i.e. the rich country) to share some portion of its wealth in order to dramatically benefit those not living within those borders.

Population refers to the new minds which will be created by the advent, as having human-

[1] Bostrom, Dafoe, Flynn

level intelligence will warrant some level of respect, and encompasses the notions of what the goals of such minds will be and what reaction their introduction to the population will cause. Protecting the interests of AI minds would be paramount in building our society off of them, after all if they present any level of consciousness even close to the of humans, there would be great moral implications in using their power to further our own goals. Should self-replicating AI be introduced to the population, there would need to be controls as to how much that self-replication, or else our society should face a big problem in terms of over-population.

Finally, the process in which the revolution was met would need to fill a couple of key requirements in order to be sustainable. First and foremost, the agency making decisions on what is permissible would need to be think on a level that isn't bound by current assumptions made by-humans-for-humans, but rather be able to process what is necessary given the current state of the revolution. Second, whatever agency is put in place to achieve this must be able to implement policies at rates that can keep up with the exponential growth of the revolution, something which our current government has demonstrated that it is less than capable of. Finally, the laws which we currently have must be able to adapt to AI and react with the understanding that the concept of AI could be ever-changing.

In contrast to the principles which would ideally be followed in the case of AI revolution, we have a compilation of the current existing suggestions and considerations from large scientific bodies and governments (AI Principles of Europe, China, US; Microsoft AI Principles; Open AI Charter) looking to put the first steps in motion when it comes to regulation. In an 'Evaluation of Guidelines'[2], researcher investigates the principles currently being recommended and what might possibly be omitted from these results. In the paper, it is assumed that companies have no real incentive to follow the guidelines put in place, only following 'guidelines' with no real consequence or preventative measures put in place that would outweigh the benefit of discovering such a technology.

Some of the principles outlined in this study overlap with the aforementioned concerns, however some additional cases more concretely include those of having privacy protection, diversity in development, the protection of those who speak out against malpractice and accountability of designers who create an AI that ends up causing severe damage to others. The study also found that the most cited issues in these studies had already been resolved, which makes sense as it makes those writing the studies (those with incentive to have less government regulation) would want to have something to show in terms of the principles that they were coming up with. It was also apparent that many of the studies were male-dominant, as many principles noted were mathematically quantifiable and logic-oriented rather than catering to more empathetic values like the welfare of robot life or what social responsibility humans owe to them. this lack of a female perspective upon creation of the guidelines seems to go hand-in-hand with the lack of guidelines concerning 'machine consciousness', which makes sense as comparatively male oriented studies likely wouldn't be as concerned with that type of issue. The article also found that there was little coverage over the potential for political abuse of the AI systems, including as noted above deep faking but also topics that have come up in recent years such as election fraud and fake news (where have we heard that in the last year). With the amount of media buzz and attention that got this time around, imagine how little trust people would have if the technology to abuse political systems improved. Also, despite one of the most common questions regarding AI in recent years having been 'what will a self-driving car do in this situation?', very little coverage of robot ethics was found in the presented guidelines.

[2] Hagendorff

Finally, there was very little talk about how to deal with any 'hidden' costs of producing such an AI, be those environmental or social. An example of a hidden social cost of AI would be the labor required to make labels to generate training data off of, as in who is making those labels and what are they getting paid/how are they being treated, as well as the amount of resources needed to just research intelligence before the advent, not to mention how much energy will be used afterward. These problems, if not addressed, are also likely to compound on each other, for example what happens when we have conscious machines that need the same energy humanity does?

Aside from all of this, we can't even guarantee that the government will have any say in how these machines are developed, as more and more often the private sector is funding the development of this technology at universities, who's to say corporations won't really end up reaping the socially-consequential benefits of these studies? In short, there are a plethora of issues that are not only being not concretely regulated, but not even discussed among the corporations that claim to be following a code of ethical AI practices. And none of this even takes into consideration what directly maliciously-minded people could do upon gaining access to the technology being developed.

One reason why companies would be ignoring ethics entirely could be due to the AI arms race as mentioned above. Namely, the US, China, and Europe all have a competitive reason to develop AI before the other, or to at least 'keep up with the pack', being that whoever isn't as caught up will be totally outclassed militarily. This effectively turns what could be a global effort in developing a universally beneficial AI into a race to see who can achieve the super intelligent standard first, and as we're well aware of ethics have always been the top priority of a competing US army.

To make bad worse, ethical guidelines that are currently in place likely aren't doing squat. A study was conducted in which software engineers were split into a control and test group, wherein the test group viewed a sample code of ethics from ACM before both groups were subjected to a series of ethical dilemmas. Depressingly, reading the code had no significant effect on the outcomes of the tests. What makes this even worse is that even if the workers on their own were likely to make ethical considerations paramount in their own answers, there is no way in hell that they would do that under the scrutiny of any business willing to hire them. The top priority of a company is to make money, and as such no competitive company in their right mind is going to slow down to make sure ethical codes are being followed if their competitors are knocking at the door of the same market share. Take Nestle for instance: they opportunize on less developed countries by selling them their own water and are typically two degrees of separation away from child labor. If you can restrict water and work children to death in the name of money, there's no way you're going to bat an eye at the potential consequences of a computer program that has the opportunity to make boatloads of money. So it should come as no surprise that there are companies using AI now to an ethical degree similar to that of a hyena's ethical degree not to attack an old gazelle. Take Facebook for instance: they actively profit off of foreign entities meddling in the corruption of US politics through the use of fake news, and all the while we have no idea what they're using their data for. In summary, it would be asinine to take large companies at their word in regard to using their AI in line with ethical purposes.

While it's easy (and not totally unfair) to blame corporations for the ethical misconduct of data usage and potential AI development, it isn't correct to assume that they aren't following legal rules in their development, for the simple truth that there are hardly any legal practices in place to punish companies for acting unethically. One reason is, of course, that government moves and passes bills through legislature slowly. As seen during Facebook's court hearing, most of the people passing laws that would keep companies in check don't have a strong grasp on the potential of AI or even what ethical guidelines should be followed while implementing laws. But even aside from the government's ability to regulate companies operating under its jurisdiction, what would the laws be? It's easy to draw up a doctrine outlining what the 'ideal' or 'most important' qualities an AI driven future would pertain to, but the fact of the matter is that there hasn't been a universally accepted concrete suggestion for what the laws should be or what the objective limits of those laws would be. This is a serious problem as given the speed at which AI technology has been developing, a law that is delivered 3 months too late might as well have been written before the stone age.

While looking ahead to the potential risks AI could present is a step in the right direction, our efforts might be better focused at current problems created by AI and what laws are being put in place to fix them, if there were any laws being put in place to fix them. There have only been a handful of AI related bills suggested to congress thus far, and out of the less than half that haven't been shut down, most of the purely entail allowing the government access into the information being used to develop the AI. A step in the right direction, for sure, but nothing that is going to result in significant impact unless built upon. That being said, it is still probably better to tackle some of the current issues surrounding AI, as fixing those will lead us to being better equipped to handle the problems of tomorrow. Many of these issues have been conveniently summarized in a report[3] which has noted these as the key current issues of AI: Lack of Algorithmic Transparency, cyber security vulnerabilities, bias and discrimination, lack of contestability, legal personhood issues, intellectual property issues, adverse effects on workers, liability and lack of accountability.

The lack of algorithmic transparency is problematic in that companies have the ability to use AI products in order to determine who gets/loses jobs, yet doesn't provide a reason as to how those workers are selected. While solutions include having whistle-blowers for when decisions are made purely algorithmically, its likely not going to do much as there are ways to develop algorithms in order to provide 'reasons' why, and it would be difficult to uphold a universal mandate to make sure companies fire uniformly fairly. Also, if accurate AI systems are developed, more likely than not humans will be able to see reasons in the employee's statistics as to why they should be fired.

Cyber security vulnerabilities are in my opinion at the forefront of what should be regulated, as much of our current commerce is dependent on the fact that information can be privatized and protected, take for instance whenever large corporations are hacked and big data breaches occur - those are problematic to say the least, not to mention the issues that would be spurred by national security intrusions. Potential solutions that were listed essentially list what developers already do to protect big machine learning systems, so the real solution would involve an experience what types of architecture have been attacked so far and in what ways could you prevent that specific attack from occurring again.

Bias and fairness have always been a problem in machine learning, already in this paper's citations there has been an example highlighting the over-representation of men in the field and no doubt this could show up in some of the programs themselves. Fortunately, there has been some research into the subject of making it a fair playing field, and as opposed to the other two there actually seem to be viable ways to test these biases within a system so in all

[3] Rodrigues

due time this could very well be a problem of the past.

Lack of contestability falls very close in line to that of algorithmic transparency, however this falls more closely to the 'how is my data being used' side of things rather than 'how is this decision being made'. The main issue here is that once a company, say Facebook, includes someone's data in one of their algorithms, there is nothing that a user can do to recollect that data. The main solution presented for this issue was to 'change the design' of the algorithms using such data, however this will likely take some great innovation on the part of the developer.

Intellectual property rights refer to the ownership of characters or media created by a machine and which human those rights belong to. Ignoring the idea that they may belong to the machine (which, according to the other articles may be a pertinent issue very soon), it seems fairly clear that the work would belong to the owner of the AI unless the machine was being employed by a separate entity during the time of creation, in which case it would belong to that entity.

Adverse effects on workers entails basically one of the core issues that has always surrounded AI, that being what happens once jobs are replaced. The proposed solutions include things like revamping education to account for work with AI or reeducation for current employees whose jobs are being taken over. That being said, with the rate at which AI develops it will more likely than not be the case of looking towards a more socialist society rather than having to keep retraining people to keep up.

Privacy and data collection issues have been in the hot seat for recent AI issues, however many of the solutions being proposed are fairly well implemented. One of which suggests that users should be made well aware what their data is being used for, and while that could be made more transparent most people don't bother looking at the term and conditions regardless. These systems of transparency and fair data usage can always be improved upon, but the basic solutions are already in effect.

Liability laws are likely some of the more straight-forward laws to be brought in, as really costs and damages can be summed and the responsibility could go to either the distributor or developer of whatever software caused the problem. Not to say that it would be an easy job, but likely would not take out of the box thinking to solve.

4 Analysis

There were many problems that AI has spawned presented in the research above, all important but some on a larger scale than others. I will use this analysis to voice my own strategy as how to best mitigate each of these problems as we spiral into a world of AI supremacy.

First, any system for solutions that we come up with will need to be able to be implemented with the condition that they can solve problems at the rate at which AI grows. The first reason for this is that no government is likely to implement a solution that involves slowing down the developmental process of AI, because as was determined in the research, any government that halts its progression of AI will be necessarily behind its competitors due to the arms race that has developed. By contrasting the ability at which governments have been able to cooperate in the past with the bounty of perfecting something like super intelligence, its laughable to think that they would slow down their progress in favor humanity, especially those who stand to be most negatively affected by an overly rapid development. Secondly, delaying the advancement of computer intelligence could adversely affect the people of today in the delaying of solutions to problems that face us today. For example, if super intelligent AI found a way to stop world hunger, delaying that progress for 5 years would result in 45 million people unnecessarily starving to death, which would be potentially more disastrous than the consequences of rushing the process. Finally, forcefully delaying the onset of AI would create a political divide between the regulatory system that manages its speed and the institutions which want to have unregulated development speed. By developing a system that allows these bodies to maintain their current speed, not only would it avoid the other two problems, but it could lay the groundwork for cooperation and healthy growth of AI for decades to come.

Also, a system regulating the development of AI would likely need to regulate the world leaders in AI equally. This would not only be a fair way to go about doing things, but it would also avoid some problems down the line which could result from holding one institution back. One of these problems, as mentioned many times, would be an imbalance in the arms race. Anything gained from regulating any but one entity's progress in AI would be nullified by the progress of the unregulated firm. Another benefit to having a worldly standard of the progression of AI with respect to ethical code would be that it would be easier to get the individual to agree should their competitors be abiding by the same conduct. Finally, having an entity with goals independent of those being chased by its subjects lends to an impartiality which would greatly benefit the common, humanitarian goals. It it wasn't dominated by US or Chinese citizens looking to give their respective countries 'the edge', it would have a more universal perspective on how the most utilitarian value could be derived from this project.

If the last paragraph didn't make it too obvious, my ideal goal would be to have a UN style conglomeration of scientists and ethicists from different countries, united in the common goal of the prosperity of humanity resulting from the changes brought upon by AI. In addition to having an unbiased view of the direction AI should be taking the world, the development control group would be able to set concrete laws in place by which participating parties must abide, with incentives for countries to join such as the collaboration of research and being able to keep a compendium of AI research up to that point, being able to report to the world the current progress of the AI situation. Also, more often than not governmental policies are what initiate conflict more so than the scientists developing the weaponry. Should countries not want to join this league, individual scientists would have the ability to do so, and given enough support the organization may be able to provide political asylum to those who wish to join but are forbade by their governments.

Obviously, this center would not be the forefront of AI research. Individual entities and countries would obviously be able to retain some level of secrecy in order to maintain that they have the competitive edge. But should enough scientists join from a diverse selection of areas, the organization would have far more data than any publicly available database is able to provide currently, and able to present governing laws in regard to AI faster than any individual government has displayed the ability to do so.

Aside from the regulatory service this organization could provide, being able to distribute cutting-edge data across the entire world would be extremely enabling to lower world countries. Especially in places like Africa that have been traditionally developmentally behind the western and eastern worlds, having a resource to catch up to the leaders in AI would be a resource like none presented thus far. Talking about the dangers of an arms race that could result from AI, one thing that could seriously mitigate that would be the knowledge that whatever progress you make could be easily replicated by the rest of the world only a short time later, not to mention the economic equality that would be presented by this.

This solution is obviously not fool-proof. For one, the organization would likely become a scapegoat for risks that have been realized by AI, which begs the issue of funding. Not only would there be an issue should the organization have to pay out due to a lack of their oversight resulting in catastrophe, but it would likely take a large financial undertaking to be set into motion. A large financial undertaking to come from where exactly? As it stands now, there would likely be no countries willing to shell out large sums of money for an organization that would effectively impede the progress in obtaining the super tool. Finally, it would likely be fairly difficult to maintain a truly impartial jury within the organization to make decisions, as if we've learned anything from the real UN, democracy only lasts as far as you can throw it, as superpowers typically have the influence to get smaller countries to vote in a way that they want. There would need to be safeguards put in place at this unification AI well wishers to make sure that one country's interests were not overly favored, which would be a problem in its own right. However, should these problems be mitigated successfully, a decisive tool for humanity's future would be created, and I believe that more benefits would come out of it than the costs taken to make it.

Should this AI-based league of unions not be realized, there is still hope for the safety and regulation of these technologies. For example, a new department branched from US legislature could be formed and the US could at least regulate itself. This would not suffer from any of the problems mentioned above, however if it were to branch solely from the US government it might take too long to spawn to be effective, as the projected date for the singularity is less than 30 years away. That being said, the notion of a space force originated with Trump and that took less than 6 years to come to fruition, so perhaps a US-centric AI regulation facility is not impossible to complete in a timely manner either.

No matter what, more research needs to be done into most of the issues listed in the research body paragraph. This is especially problematic as data is such a precious resource these days, most firms likely have it guarded to the teeth much less be willing to share it with nosy government investigators. So while the current state of affairs is grim in terms of regulation on AI, it is fairly necessary to do so lest we run the risk of facing the plethora of problems above.

5 Conclusion

In conclusion, there are many factors to be wary of as the progression of AI draws further and further into the future. One of the biggest problems is the arms race between countries, driving rapid development to not be the last ones without an AI. This drive leads to a spur of other issues being not only neglected but in many cases accelerated: take for instance the increasing rate at which jobs are being automated, which spurs on the risk of economic division, which in turn causes companies to develop even faster as to not be on the wrong foot of economic development. Another prevalent issue is that of whose shoulders responsibility falls onto when catastrophic events do occur? Take for instance when self-driving cars crash, who should ultimately pay for damages? If machines were to gain power over humans, would we even still have the ability to decide who gets punished for doing what, and to what extent? And these issues don't even come close to addressing the risks posed by threat of cyber-security threats and deep-faking and all the like.

It cannot be overstated the number and severity of risks posed to humanity by AI. An infamous quote from Spiderman reads that 'with great power comes great responsibility', and that has never been more true when it comes to dealing with programs having the potential to scale to the moon and back. It will likely take a great international effort to subvert the consequences brought upon by AI, but if we can band together to manage that, the upsides presented to us are assuredly endless. Sources:

Bostrom et al. (2020); Hagendorff (2020); Rodrigues (2020)

References

Bostrom, N., Dafoe, A., & Flynn, C. 2020, Oxford University Press: Public Policy and Superintelligent AI: A Vector Field Approach

Hagendorff, T. 2020, Minds and Machines: The Ethics of AI Ethics: An Evaluation of Guidelines

Rodrigues, R. 2020, Journal of Responsible Technology: Legal and human rights issues of AI: Gaps, challenges and vulnerabilities