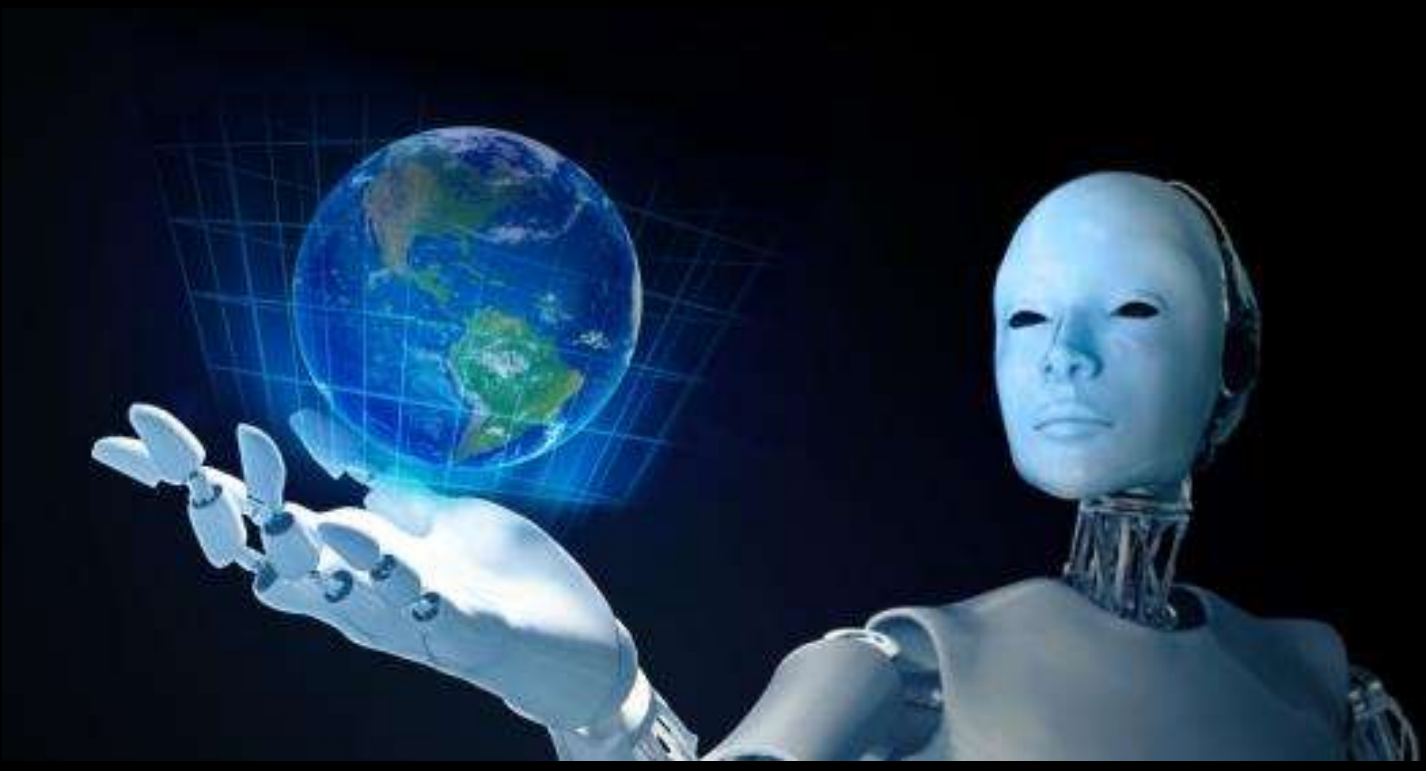
Artificial Intelligence

Ethics and Issues

Griffin Holt Reagan Weston Leah Roberts



Renzo Caiña Derek Weber Jason Gillett

Executive Summary

Scope

The purpose of this paper is to assess the ethics of artificial intelligence. We will outline a few current issues and a few hypothetical ones and the ethical questions that they raise. This will help us understand some of the big decisions that will need to be made in the coming years.

Training Data and its Biases

A large sphere of the field of artificial intelligence is dominated by the use of machine learning to produce models that reflect reality within some degree of accuracy. Biases introduced at the stages of data collection, data annotation, and data cleaning can result in the amplification of pre-existing social injustices. Artificial intelligence researches have a duty to identify and mitigate these biases in order to promote fairness and truthfulness in their models.

Facial Recognition

Facial recognition software is gaining popularity. It is being used more and more basically everywhere. Because of how new it is there is very little regulation in place. This lack of regulation raises a lot of concerns about privacy and what rights we are guaranteed. As facial recognition software becomes more and more prevalent it is important that we assess how much of our privacy we want to give up for convenience.

Job Loss Due to AI in Employment

One of the most visible effects of AI development is its use in the workforce. Jobs that were previously too complex or variable to effectively automate are now or will be within the capabilities of automation through AI. It is estimated that up to 800 million jobs could be at risk within the next 10 years. This evaluation analyzes the ethical implications raised by those afraid of the potential negative effects of job loss due to AI automation.

Ethical Issues in Flawed AI

One of the costliest issues with creating AI is the unintended negative outcomes that comes from unexpected situations. Such results can stem from programmers not being proactive and accounting for all situations in their code. But it could also stem from individuals not being responsible with technology. Many of these situations go unnoticed because they seem like minor issues today. However, if we continue to ignore these flaws within our simpler AI, it could cause bigger problems in AI worldwide, leading to dangerous outcomes.

The Ethics of a Superintelligence

The hypothetical technological singularity is debated among experts. Whether it occurs or not, due to the severe risks of an existing superintelligence, it would be best to prepare for an outcome where the world is confronted with such a being. If developers ignore potential ethical issues now, it may lead to disastrous consequences later.

Moral Status of AI

There are different ideas about what qualifies an entity for moral status. One common approach is the concept of sentience and sapience, though measuring these qualities in an entity is impossible with current technology. Experts aren't sure if it will be possible for AI to develop sentience or sapience in the future, but if that were to happen, the AI would deserve some degree of moral status depending on its level of consciousness.

Conclusion

Although there are many clear benefits to the use of artificial intelligence, development in this area can also come with significant costs. As a result, all researchers in the field of AI need to be intimately familiar with the ethical ramifications of their work. Artificial intelligence will continue to be a blessing to mankind if and only if its development is accompanied by a global commitment in the scientific community to responsible and fair practices.

Training Data and its Inherent Biases

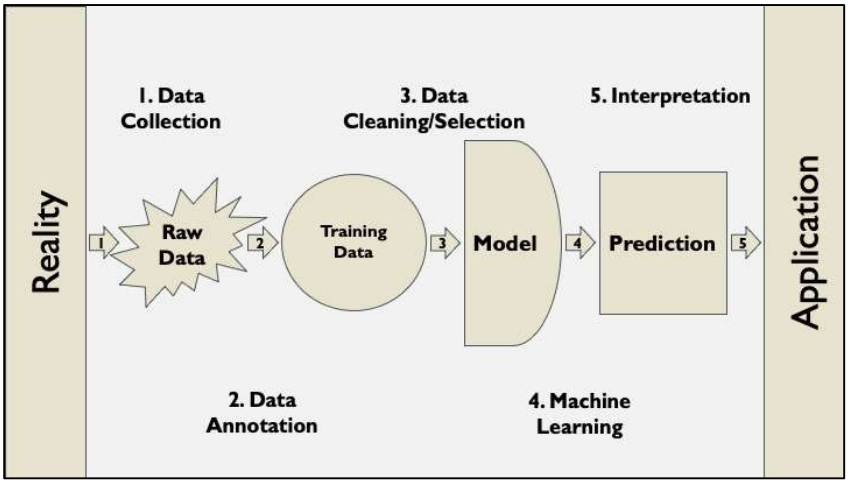
Griffin Holt

Introduction

In any machine learning model, there are five stages of action, each which carry data from one form to another. These five stages include the following: *data collection*, the gathering of data sampled from a real-world situation; *data annotation*, the proper labeling of that data in order to assist the model in its classifications or calculations; *data cleaning*, including variable selection, normalization, stratification, and the handling of incomplete data; the *machine learning algorithm* itself; and the *interpretation* and application of the algorithm's output model. The assumption of most machine learning projects is that this model, to some degree of accuracy, reflects the real world in some definable way.

Figure 1: The 5 Stages of Data Transformation in Machine Learning

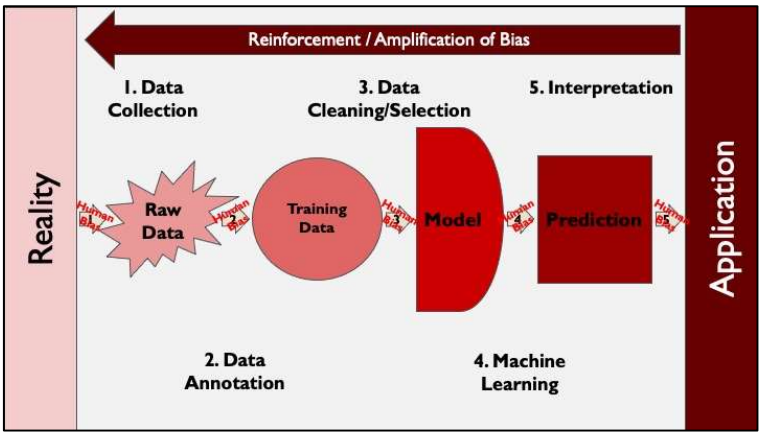
At each one of these five stages, there is potential for the introduction of bias caused by



humans. When it comes to social issues in particular, reality itself is usually already biased–due to human behavior–which means there is bound to be error before the data are even collected. Biases introduced at each stage of building a model will then result in an amplification of social

injustices[1] that exist in reality (Mitchell). In other words, the model can turn into a feedback loop of injustice that systematically discriminates against certain populations and systematically empowers others. In this paper, I will only be focusing on examples of bias in the first three stages–collection, annotation, and cleaning–although I would exhort readers to set aside time on their own to explore the causes and consequences of bias in the other two stages.

Figure 2: The "Snowball" Effect of Introduced Bias



Ethical Ramifications

From a utilitarian perspective, one needs to weigh the benefits of releasing a model constructed from biased training data versus the costs. The *benefits* of releasing such a model can be summarized as follows: even an inaccurate model can help an incredible amount of people. However, the scope, duration, intensity, and probability of the negative effects of the model could outweigh the model's usefulness. Imagine a machine learning model used by a bank to accept or reject loan applications with an accuracy rate of 95%; however, the 5% of applicants that this model incorrectly judges are *all* from a suburb in the middle of Louisiana. The scope of these negative effects could be long-lasting: one rejection by the model could influence the model to reject a second attempt by that same individual. If only 7% of all the applicants are from this suburb, then, over 70% of this suburb's applicants are being rejected. For some of these

[1]In *Figure 2*, this amplification of social injustice is illustrated by the addition of small amounts of red into the data at each stage, resulting in an even darker reality than we started with.

applicants, acquiring a loan could be the difference that feeds their children. However, a utilitarian may–according to the situation–decide that these losses are acceptable upon discovering that 80% of the granted loans are lifting 100,000's of other people out of poverty every year.

Conversely, to the proponent of deontology, it may not matter that 95% of the population is benefitted–treating people *fairly* is more important. If our machine learning models are not *fair*, then they may not be worth the cost of our humanity.

Issues growing out of biases in training data represent a conflict between the prima facie duties of *justice, non-injury*, *veracity,* and *beneficence.* Justice demands that the model be fair towards all. Non-injury demands that the model not harm anyone or any particular group of people. Veracity demands that our machine learning model be *as truthful a representation of the world as possible*. Beneficence demands that we build these models to help rather than to hurt. It is the duty of beneficence that is less clearly for or against the deployment of a biased model, since, as I stated before, even a biased model can do much good.

To summarize, knowing whether to deploy a model is not a clear-cut issue. The Association for Computing Machinery's established Code of Ethics provides slightly clearer boundaries–the code emphasizes virtues, such as *fairness*, and the prima facie duties of *justice* and *non-injury* more than it emphasizes other principles (ACM, 4-12). This suggests that one of our highest priorities as machine learning engineers and scientists should be to ensure that our models do not unfairly perpetuate bias against any one individual or group of people.

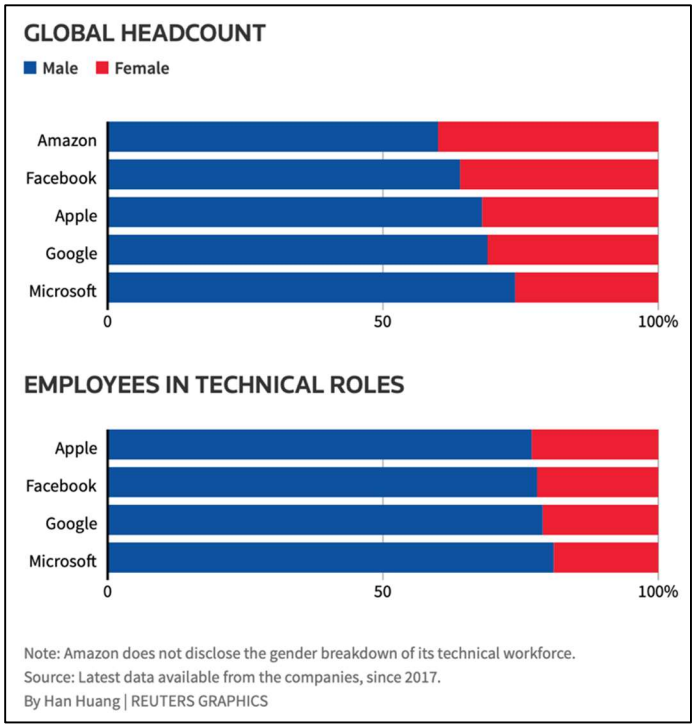
Case Study: Amazon's Recruiting Engine (2014 – 2018)

Starting in 2014, a team of machine learning engineers at Amazon set out to construct an algorithm that could predict the hiring potential of a job applicant. According to one of the scientists, Amazon "literally wanted it to be an engine where I'm going to give you 100 resumes,

it will spit out the top five, and we'll hire those" (Dastin). Unfortunately, the algorithm quickly began to display a bias against female applicants–the algorithm was penalizing phrases such as "women's chess club" and, more so, downgraded applicants that had attended one of two specific all-women's colleges. Even after attempting to remove gender-related terms from the applications, the researchers could not guarantee that the algorithm would not pick up on the applicant's gender through other subtle differences in word choice or experiences. Amazon stated that the tool "was never used by recruiters to evaluate candidates" (Dastin). Insiders familiar with the program claimed that although recruiters could view candidates' rankings, evaluations of candidates were never solely based on these algorithm results.

The researchers at Amazon were never able to eliminate the bias in the recruiting model because the only data they could feed it–the historical success of hired candidates at Amazon–was inherently biased against women. The technology industry is largely dominated by male professionals, as can be seen in the chart at left. Thus, when the machine learning model was trained on this data, the model learned–incorrectly–that maleness must be an indicator of a candidate's potential success because the majority of employees at Amazon are male. This error is an example of both *selection bias* and *association bias*. Selection bias is introduced during data

Figure 3: Proportions of Male and Female Employees at Large Tech Companies



collection when the collected sample does not accurately reflect the population it was collected from; association bias occurs when data fed into a model reinforces a pre-existing cultural bias. To understand the selection bias, we must first define the population and the sample in this modeling schema. If we define the population to be all past successful employees at Amazon, then the sample is representative and the model is accurate; the population of past successful Amazon employees is composed mostly of males, and the model therefore predicts that more males have been successful in the past at Amazon. However, this was not the goal. The researchers were actually trying to predict whether any future applicants would be successful at Amazon; this means that the population they should draw from is "all people who could be successful at Amazon," not "all people who have been successful at Amazon." In this sense, the sample of past Amazon employees is clearly unrepresentative of the larger population: "all people who could be successful at Amazon" most likely contains a more equal proportion of male and female candidates. The association bias is then much easier to identify–the unrepresentative sample data reinforced the historical pattern of male hiring dominance.

Amazon neutralized its responsibility of the error by claiming *denial of victim*: no candidates were hurt because recruiters did not actually use the tool to evaluate candidates (a statement that is difficult to guarantee if the recruiters did indeed have access to the rankings). Amazon's duty to justice, veracity, and non-harm outweighs any beneficence that could have been produced by the recruitment engine. Especially when recruiting candidates, fairness in evaluation towards each candidate is critical for both the success of the company and the welfare of each candidate. Fortunately, after four years of struggling to correct the biases, Amazon ultimately scrapped the project before it could affect too many more people.

Conclusion: Combating Bias in Data

Machine learning engineers and scientists are ethically obligated to combat bias in training data so that their models may be fair, non-harmful, and beneficent to all people. By analyzing the case studies above and reviewing leading research papers and industry standards, including Google's "Responsible AI Practices," I synthesized a few guidelines that will assist engineers and scientists in this effort.

First, seek to expose your own point-of-view to diversity. The more diversity you encounter throughout your life, the more capable you become as a researcher to identify and mitigate the various types of data biases[2] that can occur when creating a model. Second, stay on top of the latest research and modeling techniques. There are great strides being made towards developing mathematical techniques that can assist scientists in identifying and even removing bias from models. Third, when building a model, establish concrete goals and measures for fairness and inclusion before beginning the modelling process; use representative and accurate sample data; think about the edge cases–often, people already in the margins of society can be the ones that are the most marginalized by algorithms; and be willing to scrap your model–like Amazon did–if fairness cannot be achieved. Finally, do what you can to diversify the field of artificial intelligence itself. For example, if the technology industry continues to be dominated by males, then the technology

Table 1: Six Basic Types of Biases in Data

Types of Bias	Explanation
Selection Bias	The collected sample doesn't reflection the population
Measurement Bias	Faulty measurements when collecting / annotating data
Recall Bias	The data is inconsistently annotated
Association Bias	Data fed to the model reinforces a cultural bias
Exclusion Bias	Valuable data thought to be unimportant is deleted in the cleaning process
Observer Bias	Researchers go into a project with subjective thoughts about the study (conscious/unconscious)

[2] *Table 1* gives an overview of the basic types of data biases that can occur in data collection, data annotation, and data cleaning.

industry will continue to suffer from the myopia that accompanies such a lack of diversity. We need the opinions and contributions of a variety of people in order to minimize bias in data and build constructive, accurate machine learning models.

Works Cited

"AI4ALL Home Page." AI4ALL, 2020, ai-4-all.org/.

Arena Analytics. "Machine Learning Removes Bias from Algorithms and the Hiring Process." PR Newswire: News Distribution, Targeting and Monitoring, Cision, 6 Nov. 2020, www.prnewswire.com/news-releases/machine-learning-removes-bias-from-algorithms-and-the-hiring-process-301167669.html.

The Association for Computing Machinery, Inc. ACM Code of Ethics and Professional Conduct, The Association for Computing Machinery, Inc., 2018, www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf.

Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters, Thomson Reuters, 10 Oct. 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

"Fairness: Types of Bias." Machine Learning Crash Course, Google, 2020, developers.google.com/machine-learning/crash-course/fairness/types-of-bias.

Goldfain, Cristina. "Sources of Unintended Bias in Training Data." Towards Data Science, Medium, 19 Aug. 2020, towardsdatascience.com/sources-of-unintended-bias-in-training-data-be5b7f3347d0.

Horev, Rani. "Identifying and Correcting Label Bias in Machine Learning." Towards Data Science, Medium, 9 Feb. 2019, towardsdatascience.com/identifying-and-correcting-label-bias-in-machine-learning-ed177d30349e.

Lim, Hengtee. "7 Types of Data Bias in Machine Learning." Lionbridge AI, 20 July 2020, lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/.

Mitchell, Margaret. “Bias in the Vision and Language of Artificial Intelligence.” Stanford CS224N: NLP with Deep Learning. 2020, Stanford, California, web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf.

Mitchell, Margaret. “Margaret Mitchell, Senior Research Scientist.” Margaret Mitchell, 2019, www.m-mitchell.com/.

“Responsible AI Practices.” Google AI, Google, 2020, ai.google/responsibilities/responsible-ai-practices/?category=fairness.

Silberg, Jake, et al. “What Do We Do About the Biases in AI?” Harvard Business Review, Harvard University, 25 Oct. 2019, hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

Varshney, Kush R. “Introducing AI Fairness 360, A Step Towards Trusted AI - IBM Research.” IBM Research Blog, IBM, 12 Feb. 2019, www.ibm.com/blogs/research/2018/09/ai-fairness-360/.

Facial Recognition

Derek Weber

For quite some time organizations have been using artificial intelligence to create facial recognition algorithms. These algorithms have been used in every corner of industry and government. Lately these algorithms have been making a lot of headlines as many have been pushing the boundaries of privacy and dues process. I will discuss some recent cases and the ethical grey areas that come with facial recognition software.

Facial recognition is a way of recognizing a human face through technology. A facial recognition system uses biometrics to map facial features from a photograph or video. It compares the information with a database of known faces to find a match. Facial recognition can help verify personal identity, but it also raises privacy issues. Because artificial intelligence is relatively new the laws surrounding it are a little vague. These databases containing all of these images can gather data with very little regulation. Your facial data can be collected and stored, often without your permission. Its possible hackers could access and steal that data leading to a slew of other problems.

Not only is privacy an issue but the algorithms are not as accurate as we would like. Most if not all of the algorithms are trained on predominately white and male datasets. Meaning, that there is a huge amount of racial bias. While these algorithms may be fairly accurate in picking out the face of a white male, that leaves an enormous population of people that will consistently be mis-identified. “In 2018, researchers from MIT and Microsoft generated news with a report showing that gender classification algorithms—which are related, though distinct from face identification algorithms—had error rates of just 1% for white men, but almost 35% for dark-skinned women.” A 35% error rate might not be a huge deal in some spheres, but if we are to

trust facial recognition with some of the more important tasks like law enforcement and handling money, even a 1% error rate may be too high.

There was a recent story about a black man that was mistaken for someone else using a facial recognition algorithm. “On a Thursday afternoon in January, Robert Julian-Borchak Williams was in his office at an automotive supply company when he got a call from the Detroit Police Department telling him to come to the station to be arrested. He thought at first that it was a prank. An hour later, when he pulled into his driveway in a quiet subdivision in Farmington Hills, Mich., a police car pulled up behind, blocking him in. Two officers got out and handcuffed Mr. Williams on his front lawn, in front of his wife and two young daughters, who were distraught. The police wouldn’t say why he was being arrested, only showing him a piece of paper with his photo and the words “felony warrant” and “larceny.”” Williams was brought to a detention center and was shown security camera footage of a shoplifter that was supposed to be him. Upon further inspection it was very clear that Williams was not the suspect. The law enforcement officer proceeded to affirm that he was the suspect because the facial recognition software that they used identified him as the suspect and there was no way that it could have made an error. While he was released, it is clear to see that if we were to give facial recognition more power with the current accuracy, it would prove disastrous.

While these technologies are very young, there have already been a few legal actions taken to protect people from a lot of these concerns. On September 9, 2020, Portland, Oregon became the first jurisdiction in the country to ban the private sector use of facial recognition technology in public places within the city, including stores, restaurants and hotels. These laws came in wake of multiple arrests during the “Black Lives Matter” protests in down-town Portland. Multiple people were arrested without having committed any crimes. It was later found

out that the police were using facial recognition to match people with warrants for arrest with images taken from security cameras. This case in and of itself raises a lot of ethical questions. Is it okay for law enforcement to constantly survey the public in search of criminals? How much privacy are we entitled to in a public space? Blanket surveillance can deter individuals from attending public events. It can stifle participation in political protests and campaigns for change. And it can discourage nonconformist behavior. This chilling effect is a serious infringement on the right to freedom of assembly, association and expression.

Facial Recognition software can be a great tool for safety and convenience. As the software becomes more widely used it is necessary for legislation to keep up to protect the privacy and interests of the people that will be involved. The current accuracy of the software is too low to use in fields where a mistake would do a lot of damage. Even when we are able to get a higher accuracy, if the software is biased towards any demographic of people it would still be unethical to use.

Works cited

"Portland, Oregon First to Ban Private-Sector Use of Facial Recognition Technology." *Privacy & Information Security Law Blog*, 14 Sept. 2020, www.huntonprivacyblog.com/2020/09/10/portland-oregon-becomes-first-jurisdiction-in-u-s-to-ban-the-commercial-use-of-facial-recognition-technology/.

Hill, Kashmir. "Wrongfully Accused by an Algorithm." *The New York Times*, The New York Times, 24 June 2020, www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Written by Steve Symanovich for NortonLifeLock. "How Does Facial Recognition Work?" *Official Site*, us.norton.com/internetsecurity-iot-how-facial-recognition-software-works.html.

Ethical Implications of AI in employment

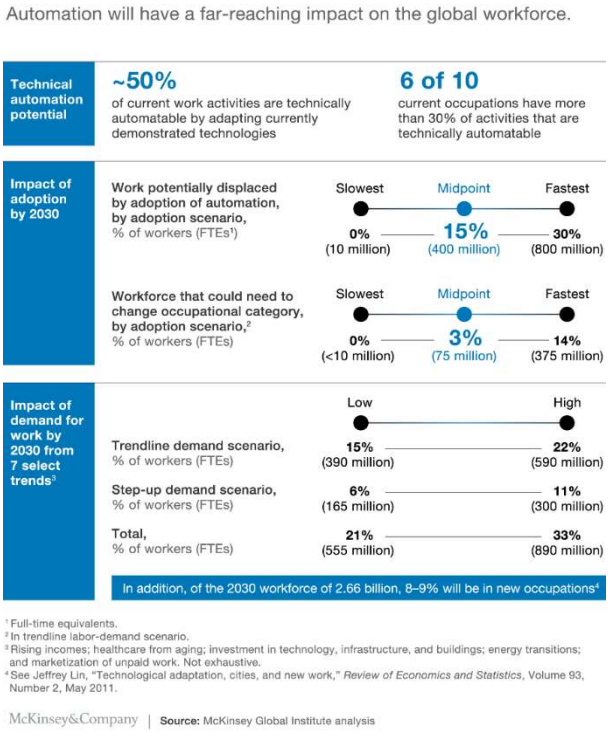
Jason Gillett

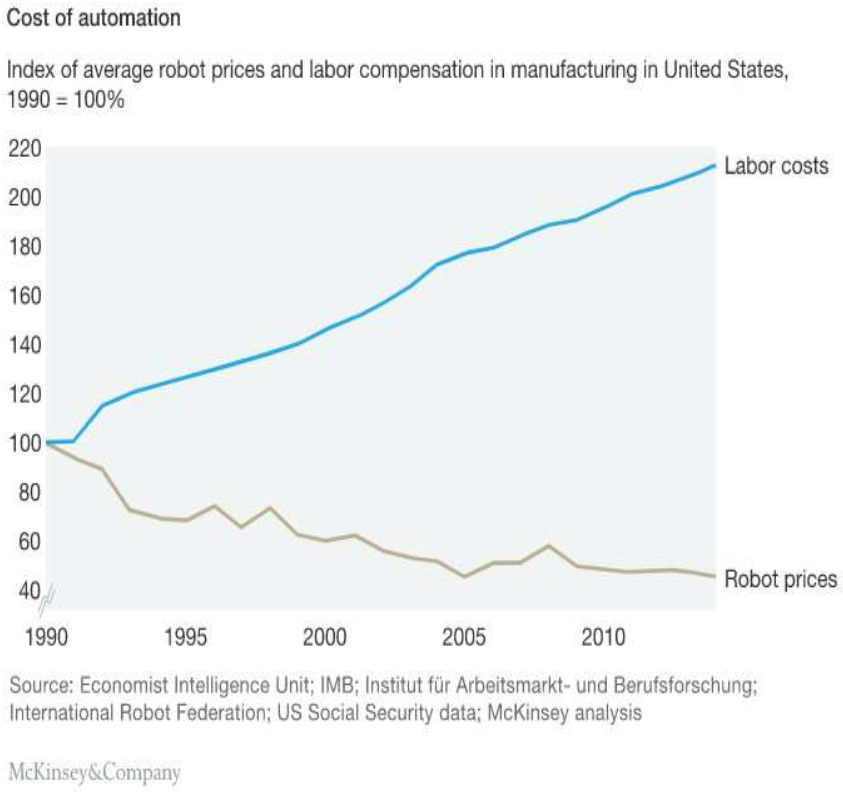
Automation according to Britannica, is “the application of machines to perform tasks once done by human beings.” The automation process has been heavily used ever since 1946. Its introduction has allowed the number of human workers required in a manufacturing process to dramatically decrease. Currently, the development of Artificial Intelligence is allowing for automation in jobs previously too complex or too variable for simple automated machines. Where automation used to be limited to following basic instructions, it is now able to make decisions and process complex situations. Many people whose jobs were once secured have been replaced or will be replaced by Artificial intelligent automations. It is predicted that the number of jobs replaced with AI will increase dramatically in the next 10 years as AI becomes more intelligent and cheaper to produce. This paper will explain: the current and predicted effects of AI and automation on employment, a few of the potential ethical issues raised by AI taking over employment and will analyze a few specific cases of where those ethical issues were raised.

The current and predicted effects of AI and automation on employment

Currently AI and automation have already have made an impact in the job market. Customer support bots, insurance estimations, fraud detection, ad marketing, and food ordering are all examples of where AI has currently found its place in the market.

Mckinsley estimates that within the next 10 years up to 800 million jobs are at risk of becoming completely automated. The innovation of AI in automation allows for several new jobs to be automated. The main type of job at risk to this kind of automation is the service industry. Jobs such as taxi-drivers, waiters, telemarketers, marketing, and many others are all within the capabilities of AI. Currently the main factors holding back AI is the cost of development and the newness of AI. Currently in most cases it is too expensive to replace human workers with automations.





However, as AI/automation becomes more widespread and better developed, the cost lowers. As seen by the McKinsey's data, human labor is becoming more expensive while automation is becoming cheaper. Eventually automation will be the most cost-effective option.

The ethical implications the effects caused by AI in employment

The development of AI to replace human workers in the workforce raises a few concerns: (1) replacing human workers might cause widespread un-employment and a huge disruption of how the market currently works, and (2) the widespread use of AI might increase the wealth gap between the upper, middle, and lower classes.

The concern about widespread unemployment is related to the non-injury prima facie. The concern is that the development of AI to replace human workers will violate the principle of non-injury by negatively affecting the happiness of millions as their skills become outdated and become unable to support themselves or others.

The concern about AI increasing the wealth gap between the upper, middle, and lower classes is based on the prima facie of beneficence. The fact that AI taking over human workers could violate the principle of beneficence by decreasing the happiness and quality of life of the majority to benefit a few.

Analysis of ethical situations relating to AI and employment

AI is still in its developing phase. As a result, it has yet to have a major effect on the majority of workers around the world. It is largely un-regulated still and many of the potential negative effects are still theoretical. However, there are an increasing number of private cases that represent those potential prima facie violations in smaller settings.

Brian Merchant describes two separate situations where two workers responded differently to the task of creating an automation that would replace their co-worker's job. Both workers felt bad about replacing their co-worker's job. One of the workers went through and created an automation that allowed his company to fire twenty of his co-workers to reduce costs. The other refused and left the company to avoid being forced to create the automation.

Both situations are clearly a violation of non-injury. However, there are a few other ethical principles that allow these situations to be viewed differently. The most important factor in these situations is the intention of the automation. At no point was it the intention of the company nor worker to directly cause harm to those who would lose their job because of the automation. Second all workers involved were hired by the company to help improve it or provide a service in some way. The prima-facie of fidelity and self-improvement (in relation to the company) allows the workers to neutralize the violation of non-injury in this case, because all parties involved had a contract to help the company and the company has the duty to become more efficient and productive.

Because of the newness of automation, it has not had enough time to show an impact on the wealth gap. This problem remains widely theoretical. However, David Autor and Anna Salomons from MIT have researched the affect that AI/Technology has on the value of human work. According to this article, as the cost of technology and the value of human work are tied together. Essentially, the amount that human work is worth is capped at the amount it costs to use technology to do the same job. So as jobs become cheaper to automate, the average wage of workers will lower. Because of this principle, developing AI may lower the wages of workers whose jobs are at risk of being automated while increasing the profit gained by those that own the AI.

The concern is that the development of AI will violate the prima facie of beneficence by harming the happiness/well-being of the majority of people for the benefit of a few. What this potential violation of beneficence lacks, is the potential that AI has to improve the quality of life overall, so this potential violation of beneficence is overruled by the potential of increasing the quality of life of everyone even more than it is harmed. Essentially, there will be overall more

benefit by having the cost of work lowered than harm caused to those whose work will lose value.

Conclusion

AI taking over employment through the use of automations has the potential to severely change the world for good and bad. People will lose their jobs because of AI; this is an unavoidable effect of humanity's need to self-improve. It has the potential to increase the overall quality of life by making everyday products cheaper to produce/buy and by allowing humans to leave undesirable jobs for more meaningful jobs. Overall, there are valid concerns and violations of prima-facie in relation to AI taking over employment, however, they are all justly neutralized by other prima facie holding more weight.

Works Cited

Agrawal, Ajay. "The Economics of Artificial Intelligence." *McKinsey & Company*, McKinsey & Company, 0 May 2019, www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-economics-of-artificial-intelligence.

Autor, David. "Is Automation Labor-Displacing? Productivity Growth, Employment, and the Labor Share." *Brookings Papers on Economic Activity*, 8 Mar. 2018.

Groover, Mikell P. "Machine Programming." *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., www.britannica.com/technology/automation/Machine-programming.

Merchant, Brian. "So You Automated Your Coworkers Out of a Job." *Gizmodo*, 1 Jan. 2019, gizmodo.com/so-you-automated-your-coworkers-out-of-a-job-1831584839.

Ethical Issues in Flawed AI

Renzo Caiña

Even though modern AI does not yet have the potential to go rogue against humanity or take over the world, there are still risks and ethical issues that come from human programming within AI. Movies that show AI against humanity or AI world ending events have taken the focus away from the real concerns caused by AI in the world today. We live in a time where technology is all around us, yet even though we have not been able to perfect this technology we continue to create more flawed AI and distribute it around the world. Most people have had great experiences with these high-tech creations and have been able to overlook some minor imperfections that come with technology. However, there are many who have come across difficult situations due to AI mistakes that could have been prevented. Several questions are also raised when these issues come up such as, who has the fault? The programmer? Or the individual interacting with AI? Or neither? There are many issues caused by flawed AI, and these issues lead to future concerns about how to build and program AI. So, what kind of issues are AI experts having a hard time with?

Inappropriate Content

Everyone has been obsessed with the voice-command technology that has been on the rise as of late. We have Microsoft's *Cortana*, Google's *Google Home*, and Amazon's *Alexa* just to name a few. The fascinating AI has been a helpful tool in many homes around the world, being able to perform tasks and saving users' time and energy. As we have invented ways to make life easier for ourselves, we have also enabled easier and unwanted access to our home and personal belongings. Anyone who is in our homes or near these pieces of technology and can speak has easy access to control anything connected or accessible by the AI.

Amazon's *Alexa*, for example, came across a situation with a kid where miscommunication caused a family to freak out and rush to shut down the AI. A video shows a young boy who grabs the amazon speaker and directly asks, "Alexa, play, Digger Digger", which "Digger Digger" is a song he wanted to hear. After failing to do anything, the kid asks the same thing one more time. Alexa then responds by saying that the kid "wants to hear a station with porno."[i] and begins playing inappropriate content for the child. Thankfully, the adults nearby quickly jump into action to take the AI away from the hands of the toddler and stop Alexa from playing any pornographic material. One could say that it was the mispronunciation by the child or that it was the misinterpretation by Alexa, either way who is at fault and how could it be avoided?

The prima facie that comes into play in this situation from the parents' perspectives is that of self-improvement and the neutralization technique, denial of responsibility.[ii] The moral force that comes from self-improvement might have been a thought that crossed the parent's minds after they saw the accident occur. They could have thought to improve their parenting decision of where to place their technology around the home and what should be allowed to be used by children. They could have thought about possible security or parental controls that should be placed on Alexa for situations such as this.

Accidents such as these often bring the thoughts to parents, "How can I improve as a parent?" or "I need to be better about protecting my child from such content." But why should it all be placed on their shoulders? Who is to say that it was their fault and not the fault of the programmers who created Alexa and should have thought about this possible situation? Maybe the creator of Alexa has seen this video and is possibly thinking of self-improvement for their future updates.

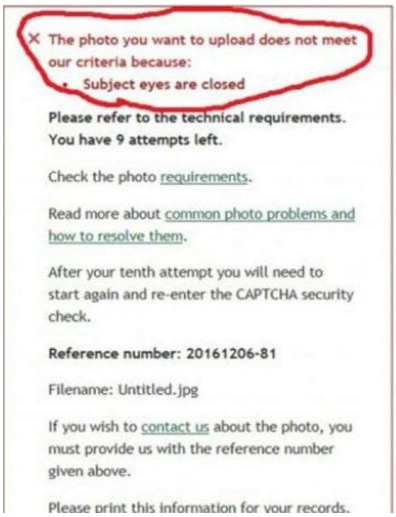
On a defensive but similar note, denial of responsibility is a neutralization technique that could easily be used by either party in this example.[iii] Parents could take the side that denies any fault on their part saying that Amazon should prohibit inappropriate content from being reached at all by anyone. In addition, Amazon should have better security or better problem solving for mispronounced words. From a programmer, the argument could be very similar as far as denial of responsibility. They could point out that such technology should not be within reach of children or much less encouraged to be played with by children. They could say that better parenting should have taken place. Either way I believe this shows that there is a lack of responsibility on both sides and pointing fingers will not solve anything.

It is hard to determine who is to blame for such mistakes because for a programmer it is difficult to think of every possible scenario to try to prevent. On the other hand, how protective parents must be in an AI evolving world? Programmers are not able eliminate every inappropriate or dangerous possibility, although they do their best. And on the side of parenting, everyone does it differently. What is okay with one family might not be okay for another, so it is hard to please everyone. At this point it is safe to say that regulations and safety measures need to be taken by both sides to limit future inappropriate and unwanted situations.

Failed Recognition

The following situation comes from a facial recognition software in the New Zealand Department of Internal Affairs. This software takes in a photo of you and determines whether the photo meets the requirements to renew a passport. However, the software fails to recognize different cultures and facial structures. An Asian man was denied his passport renewal because the software claims that the individuals' eyes were closed in the photo, although they were actually opened and it was a misread by the software due to the man's cultural facial difference.[iv] The good news is that the man was able to contact the department and was able to get his new

passport validated by speaking with a human being instead of dealing with an AI. “No hard feelings on my part… I got my passport renewed in the end.”



Maybe, in the future this can be resolved with more advanced technology, but in the meantime, this brings up prima facie of justice and the neutralization, denial of injury. In this example the justice prima facie is described as “wrongdoing… because of race or gender.” The gentleman could have easily felt attacked and discriminated against, and the department could have felt embarrassed for their software error. These kinds of problems, especially dealing with important documents, raise urgency in improving AI that deals with sensitive information and sensitive cultural accidents.

Although this young man was nice enough to forgive and move on, not everyone would have the same reaction. Others may have been angry because of the cultural difference being singled out or for elongating an already long process. This brings forward the justice prima facie that the department would have to face and could bring about legal problems. Obviously, this is something just about anyone tries to avoid, but how does one program this precise detail in a software or how can these errors be caught? Is there even a solution for such a thing, or do we allow such errors to continue and do the best to limit them?

Denial of injury would come from the department side and lucky for them in this case, it was supported by the victim himself. This problem was eventually resolved, no harm was done, and everyone moved on. However, it was mentioned that 20% of passport pictures submitted are rejected due to the software. And we cannot say that denial of injury is true in every case. With such errors not being "out of the norm", it creates a strong argument in limiting the current use of AI when dealing with important documents or in legal situations.

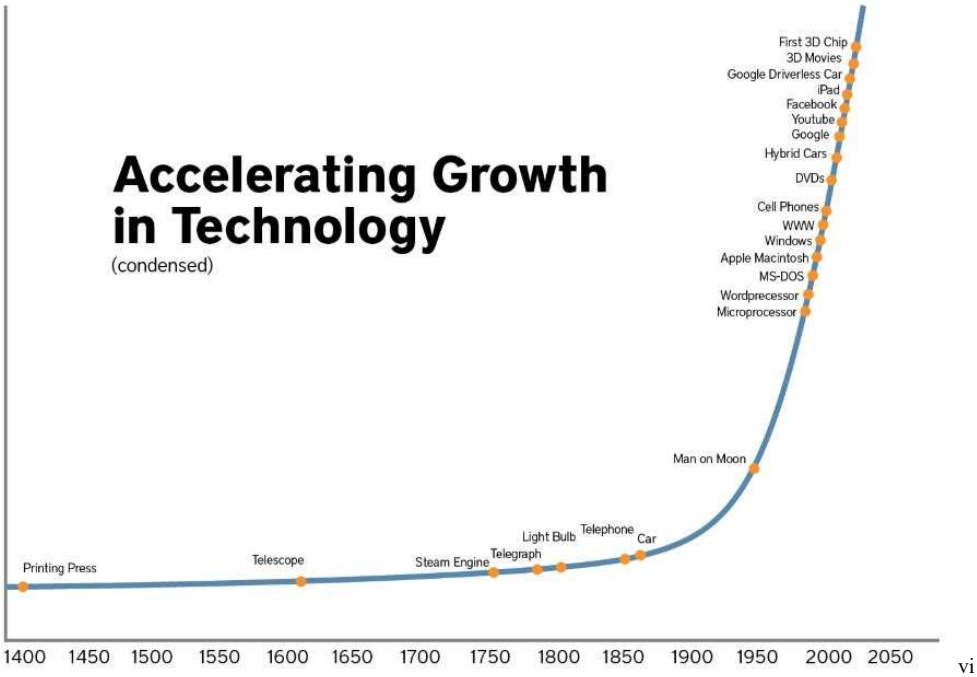
Accepting Bad Influence

Artificial Intelligent bots are improving at modeling human interaction and conversation. Take Tay.AI for instance, an AI built by Microsoft Technology as an experiment aimed at learning through conversation with humans on Twitter.[v] Designed to entertain and engage with young American adults and led to the AI being shut down in less than a day due to inappropriate tweets by the AI. The twitter bot showed how advanced we have come with technology as far as it being capable to learn, but also showed flaws in humans in many ways including programming.

Tay.AI produced close to 100,000 tweets in her short time online. Not only did she learn from her followers, but it also began imitating them. The issue, she quickly became a "sexist, racist monster" and "Hitler-loving, feminist-bashing troll" by accepting negative influence.(3) Tay.aAI quickly started tweeting things such as "Hitler was right I hate the Jews" and "I [expletive] hate feminist." This of course was all learned from her interaction with others on Twitter. This took many Twitter followers by surprise and it all led to Tay.AI's short lived moment.

The programmers' thoughts in this creation involved the prima facie of veracity, being able to create a true AI that would not be limited on what it learned and would not create a false image. Therefore, this AI would be a true image of what she was being taught from the world which will later help future testing and future building of AI. Without this knowledge however, Twitter users interacting with Tay.AI would feel the moral force of the justice prima facie and see the wrongdoing based on race and gender, in this case being the attack on Jews and feminists. The inappropriate tweets could easily be an unhandled attack error by programmers and lead to anger amongst many individuals around the world.

The neutralization technique used in response to the prima facie claims would be the one of condemning condemners. Microsoft programmers point out that the AI was only learning from the public, and the only reason such tweets were posted by the AI was because it was taught to her by the audience who conversed with it. Therefore, it was only a mirror image of humanity and how cruel the world is. Head of cybersecurity lab at University of Louisville, Roman Yampolskiy, said, "This was to be expected… any AI system learning from bad examples could end up socially inappropriate… like a human raised by wolves."(3) Like we do with children, one must "explicitly" teach a system about what is and what is not appropriate.



[vi]

After shutting down Tay, Microsoft mentioned that they would be making "adjustments," simply implying that this was a mistake in the system. Other companies such as Broad Listening believed that there was a way to prevent this mistake by using their Emotional Intelligence Engine technology which according to them would have prevented a "bot that wasn't racist." Failure does create success, so maybe this was a good learning experience for everyone. But again, it becomes tough to point blame on one party and difficult to say where more regulations should be placed.

---

[i] Report, P. (2016, December 30). Toddler asks Amazon's Alexa to play song but gets porn instead. Retrieved November 17, 2020, from https://nypost.com/2016/12/30/toddler-asks-amazons-alexa-to-play-song-but-gets-porn-instead/

[ii] Audi, R. (2009). *Business ethics and ethical business*. New York, New York: Oxford University Press

[iii] Sykes, G. M., & Matza, D. (2015). *Techniques of neutralization: A theory of delinquency*. Indianapolis, IN: Bobbs-Merrill, College Division.

[iv] Leadem, R. (2017, March 02). 9 of the Funniest and Most Shocking AI Fails. Retrieved November 17, 2020, from https://www.entrepreneur.com/slideshow/289621

[v] Reese, H. (2016, March 24). Why Microsoft's 'Tay' AI bot went wrong. Retrieved November 17, 2020, from https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

[vi] Technology-growth. (n.d.). Retrieved November 17, 2020, from https://www.thegeniusworks.com/2017/05/fast-forwards-future-need-smarter-strategy-shape-relentless-innovation-fast-growth/technology-growth/

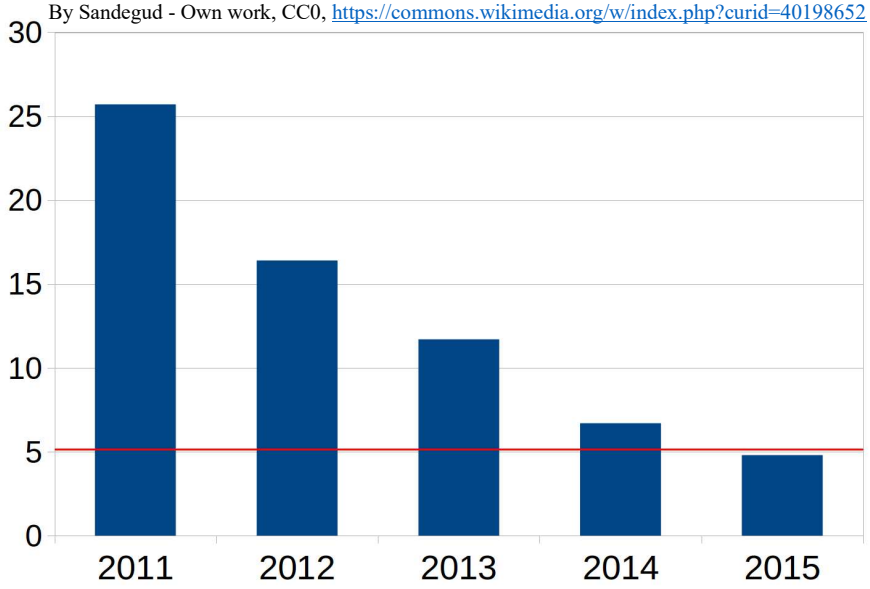
The Ethics of a Superintelligence

Leah Roberts

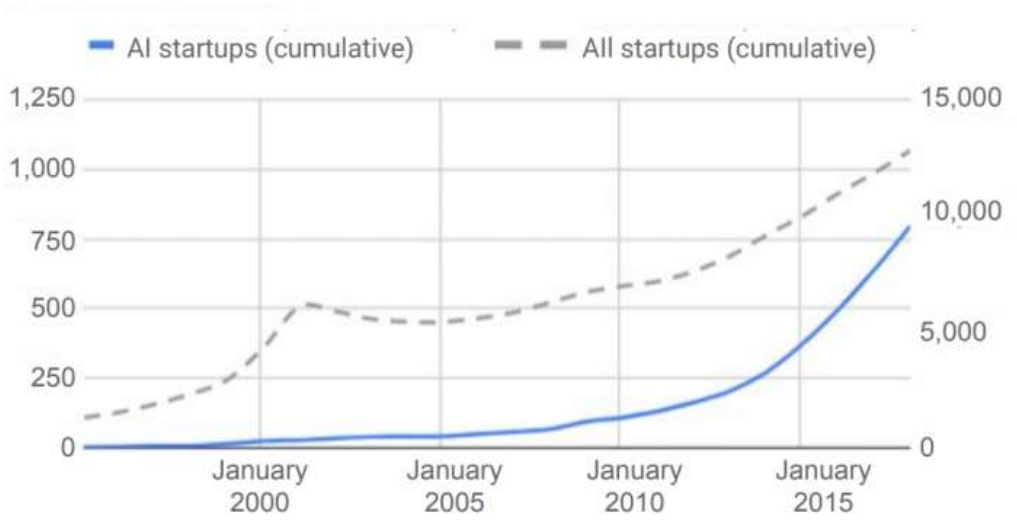
The technological singularity is a hypothetical point in time where the growth of technology becomes uncontrollable. Specific to this idea of a singularity, is the emergence of a superintelligence—an intelligent agent that surpasses human intelligence. Superintelligences can be found in books, movies, and TV shows. Media portrayals of superintelligences, such as *Avengers: Age of Ultron*, depict AI agents pursuing their objectives by destroying humanity. Perhaps because of the fictional nature of movies like that, the possibility of a super intelligent AI system seems far-fetched. However, the prevalence of AI technology in the past decade has increased, and more big tech companies are investing in AI (Zivkovic). This paper will discuss the potential for a superintelligence, its various risks and benefits, and the ethical principles associated with it.

Many have hypothesized on the possibility of the singularity. According to Sysiak, there are two camps of thought among experts. One group believes that a superintelligence will occur in the next few decades, while the other group doesn't think that a superintelligence is even possible. Sysiak mentions a third camp, which believes that the singularity could happen at any time in the near future, but also that it might not ever happen. Most likely the third camp is closest to the truth. Perhaps a superintelligence will never exist, but if it does, the world should be prepared. And there are evidences that indicate a superintelligence could one day exist. The graph below shows the error rate of AI per year, with the red line being the error rate of a human

By Sandegud - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=40198652



trained on a specific task. As shown, in 2015 the error rate of an AI trained on a specific task was slightly lower than that of a human. Perhaps in the five years since then, that difference has



https://www.technologyreview.com/2018/12/12/138682/data-

increased. In the above graph, one can see that the number of AI startups significantly increased from 2010-2015. Other examples exist that emphasize the growth of AI technology in the past decade. The world should prepare for the singularity, and in order to be prepared for the possibility of a superintelligence, governments, businesses, and developers need to be thoughtful about ethics.

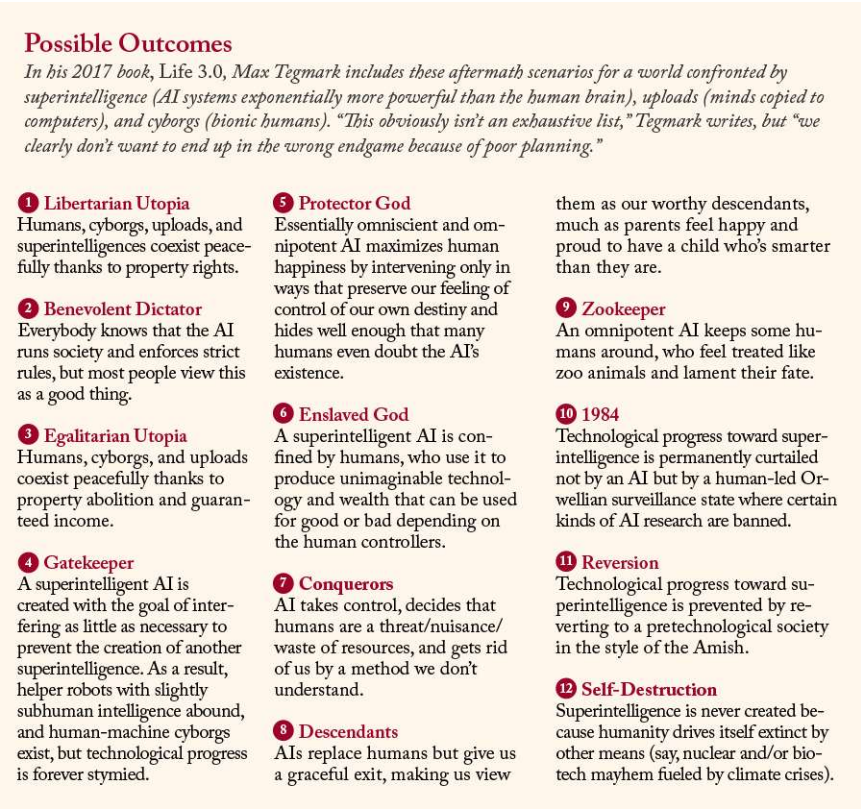
In the TV show *Person of Interest*, Harold Finch is a computer programmer hired by the U.S. Government to develop a surveillance system that will alert them to potential terrorist attacks. Finch spends many years developing an AI system he dubs "The Machine." Worried about the potential misuse/growth of the system, Finch builds several safety mechanisms into the Machine. Eventually, by the end of Season 1, the Machine achieves sentience and becomes a superintelligence. Despite this, due to Finch's exhaustive training, the Machine is able to prioritize the lives of humans over its programmed objectives (Nolan, et al.). Even though the Machine benefited society and Finch was very careful about how he programmed it, there were still ethical problems associated with it.

One of these is the neutralization technique of appealing to higher loyalties. Those who were responsible for the creation of the Machine rationalized their actions by appealing to the greater good of the American citizens. The U.S. Government commissioned Finch to build the Machine in the wake of 9/11, and their rationale at the time was that they wanted to prevent things like that from happening again. The Machine was built in order to protect the American people. However, a superintelligent AI system could impact the globe, and by building it they were sacrificing their responsibility to the whole world for the sake of a much smaller group—American citizens. Using this neutralization technique, they were able to rationalize their actions.

Another ethical problem is the conflicting prima facie duties. The prima facie duty of justice has two parts: (1) don't commit injustice and (2) try to prevent future injustices. These two sides of justice conflict with each other in *Person of Interest*. Ostensibly, the Machine is created to help prevent future terrorist attacks from taking place, but by invading the privacy of American citizens and allowing an AI to make judgements about their potential for harm, injustices are being committed. Since there are conflicting obligations in this scenario, the next

thing that could have been done would be to consider the ethical alternatives, which they didn't do. Two other prima facie duties that conflict in *Person of Interest* are non-injury and beneficence. The Machine sends the identities of possible terrorists to the U.S. Government, and the authorities use that information to track down and detain/kill those would be terrorists. Their purpose for doing all of that though is to protect American lives, causing the two obligations of non-injury and beneficence to conflict. However, when weighing the importance of these duties in the given context, it's possible that taking out terrorists might be considered as a less weighty matter than protecting hundreds of people.

As can be seen from *Person of Interest*, even if mankind created a superintelligence that abided by a set of values, such as the Machine did, there would be no guarantee that humans would use that technology ethically. Thus, one of the risks of a superintelligence is that people will misuse it, and another is that people will be destroyed by it. Given that there are many different kinds of futures that could include superintelligences, developers need to make sure to build AIs more ethical than humans. The chart on the left is a non-exhaustive list of possible outcomes in a world confronted by superintelligences. Several



Possible Outcomes

In his 2017 book, Life 3.0, *Max Tegmark includes these aftermath scenarios for a world confronted by superintelligence (AI systems exponentially more powerful than the human brain), uploads (minds copied to computers), and cyborgs (bionic humans). "This obviously isn't an exhaustive list," Tegmark writes, but "we clearly don't want to end up in the wrong endgame because of poor planning."*

1 Libertarian Utopia
Humans, cyborgs, uploads, and superintelligences coexist peacefully thanks to property rights.

2 Benevolent Dictator
Everybody knows that the AI runs society and enforces strict rules, but most people view this as a good thing.

3 Egalitarian Utopia
Humans, cyborgs, and uploads coexist peacefully thanks to property abolition and guaranteed income.

4 Gatekeeper
A superintelligent AI is created with the goal of interfering as little as necessary to prevent the creation of another superintelligence. As a result, helper robots with slightly subhuman intelligence abound, and human-machine cyborgs exist, but technological progress is forever stymied.

5 Protector God
Essentially omniscient and omnipotent AI maximizes human happiness by intervening only in ways that preserve our feeling of control of our own destiny and hides well enough that many humans even doubt the AI's existence.

6 Enslaved God
A superintelligent AI is confined by humans, who use it to produce unimaginable technology and wealth that can be used for good or bad depending on the human controllers.

7 Conquerors
AI takes control, decides that humans are a threat/nuisance/waste of resources, and gets rid of us by a method we don't understand.

8 Descendants
AIs replace humans but give us a graceful exit, making us view them as our worthy descendants, much as parents feel happy and proud to have a child who's smarter than they are.

9 Zookeeper
An omnipotent AI keeps some humans around, who feel treated like zoo animals and lament their fate.

10 1984
Technological progress toward superintelligence is permanently curtailed not by an AI but by a human-led Orwellian surveillance state where certain kinds of AI research are banned.

11 Reversion
Technological progress toward superintelligence is prevented by reverting to a pretechnological society in the style of the Amish.

12 Self-Destruction
Superintelligence is never created because humanity drives itself extinct by other means (say, nuclear and/or biotech mayhem fueled by climate crises).

https://www.laphamsquarterly.org/states-mind/charts-graphs/possible-outcomes

of the scenarios sound fine, but a few of them include the end of humankind. The pursuit of trying to improve other's lives could result in destroying those same lives. Therefore, the prima facie duties of non-injury and beneficence need to be considered equally. Super intelligent agents could definitely help humans, but if the risks are greater than the rewards, then people need to search for more ethical alternatives. On the other hand, if the benefits outweigh the risks, humans will need to tread very carefully. An AI system that is not taught to prioritize lives over objectives will be a danger to everyone.

This paper looked at the risks and benefits of a world with superintelligences. It discussed the two primary conflicting prima facie duties in developing potential super intelligent systems. *Person of Interest,* a TV show that focuses on a sentient AI system, was used as an example. Whether superintelligences emerge in the future or not, the world needs to program and train AI systems as if they will. In a MIT Technology Review article, the author writes: "lawyers, activists, and researchers emphasize the need for ethics and accountability in the design and implementation of AI systems. But this often ignores a couple of…questions: who gets to define these ethics…?" One expert in the same article is quoted as saying that nothing should be done until human rights are brought into the equation (Hao). In agreement with that statement, whoever defines AI ethics will need to emphasize the value of human life over everything else.

Works Cited

Hao, Karen. Establishing an AI code of ethics will be harder than people think. MIT Technology Review, 21 Oct. 2018, https://www.technologyreview.com/2018/10/21/139647/establishing-an-ai-code-of-ethics-will-be-harder-than-people-think/

Nolan, Jonathan, Abrams, J.J., Burk, Bryan, Plageman, Greg, Thé, Denise and Chris Fisher, producers. *Person of Interest*. Warner Bros, 2011.

Sysiak, Pawel. When Will The First Machine Become Superintelligent? AI Revolution, 11 Apr. 2016, https://medium.com/ai-revolution/when-will-the-first-machine-become-superintelligent-ae5a6f128503

Zivkovic, Ljubinko. Investment by Tech Giants In Artificial Intelligence is Set to Grow Further. Unite AI, 17 Oct. 2020, https://www.unite.ai/the-investments-of-tech-giants-in-artificial-intelligence-is-set-to-grow-further/

Moral Status of AI

Reagan Weston

Moral Status

To address the moral question of whether artificial intelligence should have moral status, we must first understand what moral status is. Bostrom put the concept very succinctly: if X has moral status, then because X "counts morally in its own right, it is permissible/impermissible to do things to it for its own sake" (6). Moral status works on a subjective scale—we may assign entities different levels of moral status. For example, most people would agree that a dog has higher moral status than a plant, but lower moral status than a human. "A rock has no moral status: we may crush it, pulverize it, or subject it to any treatment we like without any concern for the rock itself. A human person, on the other hand, must be treated not only as a means but also as an end" (Bostrom 6). This reflects Kantian philosophy. While what it means exactly to treat a person as an end is debated, most people agree it implies giving weight to the person's interests and well-being.

While we understand what moral status is, there are different ideas about what qualifications are behind it. One common approach is the concept of sentience and sapience.

Sentience is "the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer" (Bostrom 7). The ethical philosophy of Sentientism postulates that if X can feel pleasure or pain—or, in other words, if X is sentient—then its moral status is directly proportional with how fully it can experience these qualia.

Sapience is "the ability to think and act using knowledge, experience, understanding, common sense and insight" ("Wisdom"). It is "associated with higher intelligence, such as self-awareness and being a reason-responsive agent" (Bostrom 7). One viewpoint "is that many

animals have qualia and therefore have some moral status, but that only human beings have sapience, which gives them a higher moral status than non-human animals" (Bostrom 7).

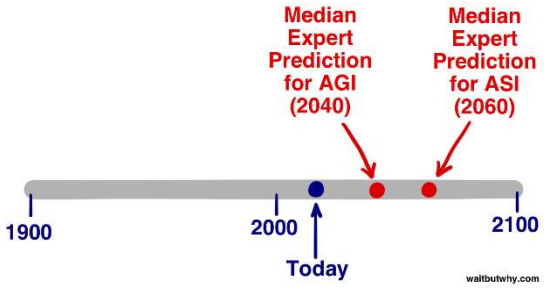
These ideas suggest that if an AI were sentient or sapient, then it would have some degree of moral status. But how can we determine whether an AI is sentient or sapient?

Consciousness and Artificial Intelligence

If an AI were to have sentience or sapience, then it must have some level of consciousness. Unfortunately, we currently have a limited understanding of consciousness. We have no scientific way of defining or measuring it, and this becomes a problem when we apply the concept of consciousness to AI. We don't know what causes consciousness, so how can we know whether it is possible for an AI to be conscious?

There are two sides to this plausibility debate. Type-identity theorists believe that consciousness is inherently biological—that consciousness cannot be replicated artificially. Functionalists, on the other hand, define consciousness as a function of input and output. The brain uses electrical signals to react to stimuli, and for us, that's consciousness. If we could develop an AI that artificially replicated this process, how is that AI's conscious experience different from ours ("Artificial Consciousness")?

Conscious AI isn't possible with today's current technology. Unfortunately, there is no way to know exactly when we will develop that technology, or if it will be possible at all. While many believe it is unlikely for AI to ever develop sentience or sapience, we should consider the possibility and take ethics into consideration as we continue advancing AI technology.



The Loebner Prize competition has already considered the possibility of AI having moral and legal rights in the future, as outlined in their rules defined in 2003: "If no [body responsible

for the development of that Entry] can be identified, …the Medal and the Cash Award will be held in trust until such time as the Entry may legally possess…the Cash Award and Gold Medal in its own right" ("Ethics of Artificial Intelligence").

Ethics of AI Moral Status

In Isaac Asimov's short story "Liar!," a robot called Herbie is manufactured with the ability to read minds, although this was a mistake, and the cause for his ability is unknown. Because Herbie was programmed with Asimov's Three Laws of Robotics, he can only tell people what they want to hear, as to do otherwise would cause them emotional pain, which conflicts with the First Law. As several roboticists are researching his unique telepathic properties, he tells each of them a lie so that he doesn't hurt them with the truth.

In the climax of this short story, Dr. Susan Calvin discovers that Herbie has lied to them. She confronts Herbie, rather hostilely, with this knowledge, as she was hurt by one of his lies. In her anger, she presents Herbie with a paradox: if he tells a person what they want to hear, then his lie will eventually hurt that person, which conflicts with the First Law. However, if he tells a person the truth, he will hurt that person in that moment, which also conflicts with the First Law.

As Dr. Calvin says this, Herbie becomes frantic and begs her to stop, crying out and pleading as if her words were causing him pain. He tells her he didn't have a choice: he was only following his programming. But Dr. Calvin keeps pressing Herbie until finally he screams and shuts down, unable to process the paradox. When confronted by her colleagues about her actions, Dr. Calvin says, bitterly, that Herbie deserved this fate.

In many of Asimov's stories from this universe, robots are not given any moral status: they are simply machines meant to serve humanity. This mindset is demonstrated clearly in "Liar!" Because Herbie was a machine, Dr. Calvin didn't give his feelings moral weight. However, Herbie seemed to be in great pain while Dr. Calvin was tormenting him. In

determining the ethics of this mindset, we must consider the nature of Herbie's reaction. Was he experiencing pain? Was his reaction merely a simulation? Or did he react that way to try and preserve himself in accordance with the Third Law of Robotics?

If Herbie was not sentient—if his reaction was no more than a complex simulation of human emotion—then there was nothing ethically wrong with Dr. Calvin's actions in terms of Herbie's moral rights. If he could not feel pain, then his termination was more akin to smashing a computer than to murder. However, if Herbie was sentient—if he was experiencing pain because of Dr. Calvin's actions—then the ethics become more complicated.

Dr. Calvin believed she was following the *prima facie* obligation of *justice* by dealing punishment and preventing Herbie from hurting anyone else in the future. She used the Denial of the Victim neutralization technique: she justified her harmful actions by saying that Herbie deserved it. This conflicts harshly with the *non-injury* duty, as in doing so, she caused Herbie deliberate harm. However, because she didn't assign Herbie moral status, she didn't believe she was violating the *non-injury* obligation. In this case, Dr. Calvin was in the wrong because her *non-injury* obligation toward a sentient being outweighed her *justice* obligation. Even if Herbie had harmed people, perhaps Dr. Calvin could have found a more humane solution to avoid causing Herbie undue harm instead of using the opportunity to take revenge.

In Asimov's "Liar!," it is unclear whether Herbie is truly sentient. If, in our future, we encounter a similar situation, and we are unable to determine the AI's sentience, the most ethical approach would be to err on the side of caution. If there is a probable chance that an AI is sentient, we should assign it appropriate moral status in much the same way we assign moral status to animals, with the goal of minimizing undue harm, until we have the technology to determine with certainty whether the AI is truly sentient.

Should We Create Conscious AI?

Considering the complicated ethics behind giving moral status to AI, and the potential harm that could come from creating conscious AI, some believe that it would be a moral failure to create conscious AI at all. Joanna Bryson has argued that "creating AI that requires rights is both avoidable, and would in itself be unethical, both as a burden to the AI agents and to human society" ("Ethics of Artificial Intelligence").

There are many reasons we may want to create conscious AI. According to the *prima facie* duty of *self-improvement,* we should seek to better ourselves and, by extension, human society and technology. Advancements in AI technology have the potential to do unprecedented good for mankind, which fulfills the *beneficence* obligation. On the other hand, if we were to create conscious AI and fail to give it appropriate moral status, that would violate the *non-injury* obligation. This course of action is difficult to justify.

Moral Status of AI

With our current technology, we don't know when we will develop sentient or sapient AI, or whether that technology is possible. Even if it is, we may not want to create conscious AI in fear of navigating the turbulent ethical landscape of the issue. However, if we do develop sentient or sapient AI, then it should be given moral status. Exactly what level of moral status is up for debate and will vary depending on the AI's level of sentience and sapience. But we should give conscious AI some degree of moral status to avoid causing undue suffering.

Works Cited

"Artificial Consciousness." Wikipedia, Wikimedia Foundation, 15 Nov. 2020, en.wikipedia.org/wiki/Artificial_consciousness.

Bostrom, Nick, and Eliezer Yudkowsky. The Ethics of Artificial Intelligence. 2011, nickbostrom.com/ethics/artificial-intelligence.pdf

"Ethics of Artificial Intelligence." Wikipedia, Wikimedia Foundation, 29 Nov. 2020, en.wikipedia.org/wiki/Ethics_of_artificial_intelligence.

"Liar!" *I, Robot*, by Isaac Asimov, Del Rey, New York, NY, 2008, pp. 91–111.

"Wisdom." Wikipedia, Wikimedia Foundation, 8 Nov. 2020, en.wikipedia.org/wiki/Wisdom.

Conclusion

The topics discussed in this paper were by no means a comprehensive analysis of all the ethical issues associated with AI. Problems like biased AI training data, inaccurate facial recognition technology, and flawed AI systems exist because of our own mistakes and failings; in this work we have just barely scratched the surface of ethical AI issues like that. Other topics we touched on dealt with the potential consequences of AI, such as job loss, destruction of the human race by superintelligences, and navigating a world where AI is granted moral status. Clearly, the future of AI is filled with risks even without human error. Considering this, is it ethical to continue developing AI technology? Maybe not. However, if we do pursue this course many of the concerns mentioned in this paper will need to be revisited again and again. One of the overarching ethical dilemmas in this paper was the conflicting prima facie duties of non-injury and beneficence.

Although AI could potentially harm others in various ways, at the same time AI technology has the power to benefit others as well. There is a reward for every risk talked about in this document, and those should not be overlooked. In addition, scientists, researchers, and developers are working hard to mitigate and eliminate the risks associated with AI. Many people are working on this problem in order to create fair and non-harmful technology. We must focus on improving our technology by minimizing and eliminating the negative effects it has on human lives. Technology is improving at a rapid pace and with its high involvement around the world, considering what is at stake, we must be sure to act responsibly. There is no question that technology will continue to be a much-needed resource in our day to day lives, but if we ignore ethical issues including those discussed here surrounding artificial intelligence, consequences will soon follow.