Ethics in Machine Learning

Brian Detweiler

October 30, 2017

Abstract

We are at the forefront of truly useful Machine Learning and Artificial Intelligence (AI) and it promises to be a revolution unlike any mankind has ever seen. As with any technological revolution, there are bound to be moral questions of good and bad, just and injust. How should we as engineers and researchers be preparing for some of these upcoming challenges? How should we be addressing existing concerns? First we will look at some of the problems we face today, from unintended consequences to biased data, safety concerns, and AI built with the intention of cuasing intentional harm, government surveillance, and blue collar job displacement. We conclude with suggestions on how we can instill ethics in our institutions and why it is necessary to include it in the discussion around AI going forward.

Rise of the Machines

> "A species had been armed too heavily - by spirit made almighty without, but equally a menace to its own well-being. Its weapon was like a sword without hilt or plate, a two-edged blade cleaving everything; but he who is to wield it must grasp the blade and turn the one edge toward himself."
> *The Last Messiah* (Tangenes 2004)

Machine learning, to the layperson, is a class of algorithms that have the ability to adjust behavior, or "learn", in response to new data. It is a subset of artificial intelligence that grew out of computer science and statistics and is often generally referred to as artificial intelligence, or AI (in this paper, we use the terms interchangeably).

Due to some of the major headlines AI has received in recent years, much of the attention has gone to the artificial neural networks behind self-driving cars, facial recognition, and winning complex games such as Go and Texas Hold'em. Machine learning can also be as simple as a least squares linear regression model, something taught in most introductory statistics courses.

From the most complex algorithms to the simplest, there is an inherent potential for harm that we must grapple with when engaging in the creation and use of AI.

Brilliant minds such as Stephen Hawking and Elon Musk warn of dystopian futures where robots are autonomously killing humans. (Vincent 2017, Cellan-Jones (2016)) Some estimate that future to be a real possibility as soon as the year 2045 (Nesbit 2015). That level of technology could come much sooner, much later, or never fully see its way to fruition, but there are very real ethical concerns that we need to be considering today, and we are very much behind the curve.

Unintended Consequences

Much of the conversation around concerns over AI is dominated by future *what if* scenarios. But there are more pressing issues that we should be concerned about today. Over the past decade, with the help of processor technology and cloud computing, machine learning algorithms have graduated from novelty applications to integral parts of our lives. One need look no further than the miniature computer in our pockets that does everything from telling us which routes to take based on real-time traffic data, to seeing social media posts that are automatically curated to our interests, to hailing rides with strangers.

Every swipe, every click, and depending on our privacy settings, even our geographic movement is generating data that is being stored away for an indeterminate period of time on data stores around the globe on which companies may train machine learning models to either uncover existing patterns in our behavior, or to predict some future outcome, be it continuous or categorical.

Uncovering patterns usually falls under the umbrella of *unsupervised learning*. Mathematically speaking, there is no response variable y, so we begin only with a set of input variables, X, on which we hope to uncover some correlation or categorization. This is typically called *data mining*, as we are looking for some pattern in the data without knowing exactly what it might be.

Prediction resides under the umbrella of *supervised learning*, in which we have a matrix of inputs X and for each row, we have a response variable, y. We then split this into a *training set* and a *test set*, and it allows us to train a model using examples of correct answers, and then test that model with examples that our model has not yet seen, but to which we know the answers. We then use this to generate a score statistic, and when we are satisfied with our model, we can begin to use it on new data to generate a predicted value, $\hat{y}$.

While all of this sounds fairly straightforward, there are "gotchas" at nearly every point along the way, and it all begins with the quality of the data. Usually when we train models, we are in control of the data, but some models take lots and lots of data to train and it may be easier to crowd source that data. As you may have guessed, there are problems with this approach.

Tay was an experiment by Microsoft researchers who wanted to train a Twitter bot on the nuances of teen slang. This included the use of emojis, drawn out words using repeated letters, and colloquialisms used by young Internet-savvy kids. Unfortunately, word of the experiment got out to *4Chan* and *8Chan*, anonymous forum boards simultaneously responsible for some of the most popular memes on the Internet as well as some of the most vile hate speech. Users of the boards coordinated an attack that sought to train *Tay* instead, on Nazi and anti-feminist rhetoric. (Chiel 2016)

The trolls were ultimately successful and the *Tay* experiment was pulled after just sixteen hours. Much of the blame can rightfully be directed at the trolls for putting an abrupt end to an otherwise interesting experiment. But ultimately, the responsibility rests on the shoulders of the researchers. Anyone who has ever programmed a web form knows that the first rule is to never trust the users. By allowing *Tay* to learn from anonymous users without filtering out racist or sexist sentiment, the researchers gave up control of their model and put it in the hands of the public. This is a recipe for disaster every time. (West 2016, Rodriguez (2016))

Biased Data

Even when we completely control the data we feed into our model, problems can still arise. Data Scientists often spend more time simply wrangling and cleaning data as they do training and assessing models. In the process of wrangling, cleaning, and validating data, it is easy to overlook other aspects of the data.

It is important to know and understand the data generating process. Where did the data come from and how was it generated? We may or may not have control over the data generating process. Whether we do or not, it is important to understand what is actually in our dataset. Following the "garbage in, garbage out" principle, even the best algorithms will perform terribly or unpredictably if trained on bad or inadequate data.

For instance, let's say we wanted to train a classifier to recognize handwritten digits from zero through nine. We have a large set of labeled images of handwritten digits. Let's assume we have 1,000 examples of each digit, zero through eight, but only ten examples of "9". It is then quite likely, than when our classifier encounters an oddly-written "9", that it misclassifies it as an "8" or a "7", or even a "1".

This is an example of biased training data, and it can manifest itself in some pretty embarrassing ways. In June of 2015, Google fell prey to biased data when its new photo tagging algorithm mistakenly misclassified a black couple as "Gorillas". Twitter user Jacky Alcine tweeted the faux pas at Google, and one of their engineers responded quickly and apologetically, stating that the result was unacceptable. (Grush 2015)

If the result was not intentional, it was likely an issue of biased training data. When it comes to humans, even those with the best intentions still have an implicit bias that can affect their ability to remain impartial. (Stecher 2017)

We should not be faulted for our implicit biases, as they are part of our evolutionary human nature, but we can and should be faulted for not correcting for them in our model selection and training. When our predictions potentially involve humans in some scope, lack of diversity in training data is bound to be a liability.

Imagine if Spotify's founders only listened to heavy metal and thus only trained Pandora's recommendation algorithm on heavy metal because they didn't think to account for other people's tastes and interests. This sounds ridiculous of course, but it is an extreme example of implicit bias. When training algorithms on human-related data, we must make sure to look outside of our own spheres of interest and assumptions, and remember to be inclusive of all types of people. If done correctly, this will result in more robust models with a lower error rate, and ultimately happier users.

Safety Concerns

Naturally, safety is at the forefront of concerns for machine learning engineers. The more useful the AI, the more potentially harmful it can be if it fails, or makes an incorrect decision. Of course, much of AI is being developed *because of* safety concerns. One example is Positive Train Control:

> PTC is a set of highly advanced technologies designed to make freight rail transportation — already one of the safest U.S. industries — even safer by automatically stopping a train before certain types of accidents occur. (n.d.)

The federally mandated PTC was enacted in 2008 after a head-on collision between a freight train and a passenger train, after the conductor of the passenger train had missed a signal while texting a friend. The PTC systems will give humans a chance to take action, but when they fail to do so, the system takes over and stops the train.

The electric car maker, Tesla has a similar crash prediction system. There are heart stopping YouTube videos of the crash prediction system alerting drivers before the human eye can even detect a problem. Many times the drivers are able to narrowly avoid an accident that would otherwise have been inevitable.

PTC and Tesla's crash prediction system are examples of AI designed to aid humans in an activity they would normally be preforming alone. Such technology is not very dangerous in and of itself, because it is simply stopping a dangerous action or alerting the user to a problem, which, in its absence, would have been more difficult or perhaps impossible to detect by humans.

As we begin to trust and rely on AI more and more, however, problems begin to surface. Some AI may be low stakes enough that we could afford a critical failure. A Roomba, for instance, may pose no greater harm than scaring the cat or scratching the floor. A one-off purchase of a shovel on Amazon may result in recommendations of gardening instruments for weeks to come. A guest who watched a show you hate on Netflix might have negatively altered your recommendations for a while.

None of these inconveniences are as serious as something like Tesla's autopilot (autonomous mode). When we ratchet up the physics and add human lives into the mix, the stakes increase dramatically. Tesla's autopilot makes the car semi-self driving. There should be extra emphasis on the *semi-*, as the car is not fully autonomous. Tesla warns its drivers of the semi-autonomous nature of their cars, and they warn that autopilot should only be used as a sort of advanced cruise control to take some of the stress out of driving, not as a chauffeur. ("Testing Tesla's Autopilot System at 70mph" 2015)

Autopilot and Positive Train Control are still in their beta stages and are likely to get better over time, but one could imagine train engineers feeling safer with PTC, and thus being less alert on the tracks, or drivers putting more and more trust into a crash prediction system, so much so that they become complacent enough to stop watching the road.

This is exactly what happened to a former Navy SEAL in Florida. Joshua Brown died when his Tesla crashed into a tractor trailer, passing underneath the trailer and into a pole. It was later found that the autopilot sensors failed to notice the white semi-trailer against a white overcast sky, and thus the Tesla failed to brake. However, it was also discovered that Brown was watching a Harry Potter movie on an aftermarket portable DVD player while trusting Tesla's autopilot entirely. The system warned him as many as seven times, which he ignored. (Reuters 2017)

Dire consequences are bound to occur as users pour more and more trust into AI systems. It is quite possible that the systems are safer than we as humans could ever be. It is also quite possible that these systems are very new and still prone to error. Trust will be earned with time and demonstrated safety.

Systems are being developed so fast though, that the urgency to rush systems to market could be superseding our desire to deliver safe products. Tesla expects to demonstrate fully autonomous capability in its vehicles by the end of 2017, and deliver them to market by the end of 2019.

Not only must we, as engineers, integrate safety features into our algorithms, but we must make sure our users know the risks and how to operate our products safely.

Intentional Harm

You may be familiar with Isaac Asimov's Three Laws of Robotics. They are:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such a protection does not conflict with the First or Second Law. (Asimov 1950)

Asimov published the laws in 1942 in a short story called *Runaround*.

These laws have been the subject of much philosophical debate over the years. Yet the United States has yet to take much formal action on AI. Tesla founder is one of several prominent figures at the forefront of AI that is worried about the situation.

> "I keep sounding the alarm bell, but until people see robots going down the street killing people, they don't know how to react, because it seems so ethereal." Elon Musk, CEO Tesla Inc (Vincent 2017)

Robots that cause harm are already here. Drones have been used in warfare since Afghanistan, but up to now, they have been piloted by humans remotely. Autopilot systems exist today that automate everything but taxiing and takeoff in commercial and military aircraft, but it would not be inconceivable to automate the entire process, and even automatically select and fire upon targets. (Reuters 2017)

Neural networks trained on hundreds of thousands of manned and unmanned aircraft mission data - much like Tesla's autonomous mode is trained on driving data - could potentially build a profile of enemy targets. One could see use in autonomous armed drones first acting as scouts and reporting back to a central command when a target is found allowing a human to take over at that point. Scouting automation would greatly improve efficiency of the pilots and allow them to save their concentration for when they need it to make critical decisions and eventually pull the trigger.

We humans are never content with our achievements though, and it is unlikely that it would end there, particularly given that we have the technology to do so. Autonomous military aircraft could conceivably perform every task from takeoff to target acquisition and execution to landing. Even refueling (or perhaps recharging) and reloading of ammunition could be an automated process.

A type of autonomous factory robot in use today that act as "movable shelves" have about an eight hour charge. These robots navigate the factory floor moving items from point A to point B. When their battery starts to run low, they simply return themselves to their charging stations for a power up. (Staff 2012)

"Blue force tracking" is a technology used by the military today uses GPS-enabled devices on all personnel to provide Tactical Operations Centers and central command with a single Google Earth-like view of all friendly assets. (Pomerleau 2015) Tapping into this information could give autonomous drones a clear classification of friendly forces, and allow it to focus solely on civilians and potential enemy forces.

Now, one can begin to picture a continuous fleet of flying armed scouts that patrol the skies of war zones looking for enemy combatants or terrorists. How long before such a system is given full authority to classify and act on targets?

This is uncharted territory. In our short history of heretofore unimagined achievements, very few rival the ethical grey areas of autonomous warfare. Nuclear war and weaponizing spacecraft are the closest situations that come to mind. Space was was regulated and treaties were agreed upon before it could become a major issue, but the same cannot be said for atomic weapons. They were developed and used before any kind of treaties could be agreed upon and hundreds of thousands of people died.

The United States, the first nation to develop the technology, was at war with Japan, and the bomb played a major role in ending the war. But soon after, the Soviet Union also had the bomb and the two nations were locked and loaded at each other on track to end the world. Americans in favor of autonomous weapons may not be considering that the United States would assuredly not be the only nation with this technology.

Indeed, Russian weapons maker Kalashnikov, maker of the ubiquitous AK-47, is already slated to deliver fully autonomous drones that can perform target acquisition and decision making. It is easy to see how an autonomous arms race could return us to the days of nuclear proliferation. (Gilbert 2017)

In 2015, a group of more than 1,000 AI scientists and researchers including Stephen Hawking and Elon Musk signed an open letter warning of such an arms race.

> "If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow." (Various 2015)

We have to decide if such a world is one we are willing to live in. It's easy to imagine a world in which robots just fight each other, but that is a naive assumption. Autonomous robots will kill humans - it is a matter of time at this point. Multinational treaties could prevent this, but with the current state of world affairs, governments don't seem to have much interest in slowing down autonomous weapons when they clearly provide such a huge short-term advantage to the owner. A legal solution is likely to be further off than an actual deployable autonomous drone, so we may be an in arms race situation before we have a chance to think much about it.

Automation Killing Jobs

To complicate matters, we are in the midst of the most significant technological revolution in history. While previous technological revolutions like steam power and the internal combustion engine put some segments of workers out of jobs, they also created more interesting jobs. This may not be the case for the AI revolution.

Kai-Fu Lee, founder of Sinovation Ventures and a leading voice on AI, says that AI is potentially replacing everything. He believes the AI revolution will dwarf all other human technology innovations combined. ("Artificial Intelligence Will Replace Half of All Jobs in the Next Decade, Says Widely Followed Technologist" 2017) While it is impossible to stop the march of progress, we should also be approaching this revolution with care. If job sectors are disrupted too suddenly and people are forced out of work with no alternatives, we could find ourselves in the midst of an economic crisis and potentially a middle-class revolt.

Few politicians understand the coming threat of automation. Former President Barack Obama sees automation and the loss of blue collar jobs as a great divider and the driver behind the cynicism that seems to dominate

modern populist political culture.

> "The next wave of economic dislocations won't come from overseas. It will come from the relentless pace of automation that makes a lot of good, middle-class jobs obsolete." Barack Obama, Former U.S. President (Miller 2017)

Obama's proposed solutions of more education in the form of early pre-school, free two-year community college, and pushing computer science in high schools seems better than nothing, but yet falls short of a bonafide solution. Making everyone code-literate can't hurt, but coding isn't for everyone. And those who don't want to code and don't have the option of a blue collar job are left in the wings. One of the industries that will be most automation-proof are service industry jobs, which include mental and physical healthcare professionals and restaurant and bar staff. (Dodgson 2017) Yet wages have been stagnant for years and politicians don't seem to want to address the issue.

Nebraska Senator Ben Sasse is one of the few congressmen speaking openly, not only about the impending consequences of automation, but also about our lack of a solution.

> "[Trade] is a much smaller topic than Artificial Intelligence. ... We can talk about a specific factory moving from Ohio or Indiana to Mexico, and the jobs that might be saved or lost in a move like that, but the much bigger long-term factor is that each of those factories has so many fewer workers. We're talking about 7% of the U.S. workforce now working in industrial jobs. And we're not wrestling with any of those questions and neither political party has an answer. [People who think that] this town is made up of a whole bunch of geniuses who can look into their crystal ball and know how to centrally plan 2030? We don't." Ben Sasse, Senator (R-NE) (Sasse 2016)

While it is encouraging to hear a sitting senator talk about this very real threat to a middle class workforce, it is at the same time discouraging to see it followed up with a mere shrug. The AI community and politicians need to work together to come up with an agreement on what is happening, and a game plan for how to deal with it. We simply cannot allow ourselves to be thrust into a world where nearly 50% of all jobs become automated and an under-educated workforce becomes unemployable because they were told their factory jobs would be saved. (Berger 2016)

Big Brother Using AI

In *1984*, George Orwell envisioned a terrifying future of absolute obedience to a government ensured by around-the-clock surveillance. Orwell had based his novel on the oppressive regime of the Soviet Union, and added some technological embellishments using his literary license. In the actual year 1984, it seemed far fetched. But now, thanks to AI, we have exceeded that capability by leaps and bounds.

Facial recognition and target tracking is a very real and increasingly practical technology that is being deployed by governments all over the world. Most notably in China, where 20 million CCTV cameras (ominously named *Sky Net*) have been installed with AI that can identify a person's age, gender, and clothes. It can identify vehicles as well. (Lo 2017) The stated use case is to help police track down criminals, but as many already consider China an oppressive state with a bad record of human rights, it is clear how such a powerful network of monitors could be used to control a population.

Russia appears to be moving ahead with surveillance too. Moscow is AI to it's sprawling network of 170,000 CCTV cameras, saying it will help them catch criminals. In what is becoming a familiar problem-solution pattern in the era of Big Data, the police have a treasure trove of video footage but cannot process it by hand. But with the help of machine learning, facial recognition software can be run on the footage in a fraction of the time and persons of interest can be flagged. The cost to deploy the technology to all 170,000 cameras was admittedly too expensive, so it is being deployed selectively, but technology costs will continue to drop and there is no reason to assume that a government would be unwilling to spend the money if they had a reasonable incentive. (Khrennikov 2017)

Here in the United States, artificial neural nets are being used for another kind of monitoring. The Edward Snowden document dump revealed that the NSA employs neural nets for speech-to-text translation, which

then allows for natural language processing (NLP) algorithms to analyze spoken words and flag conversations that either seem similar to other individuals the NSA is interested in, or very different from the population.

> "When the NSA identifies someone as 'interesting' based on contemporary NLP [Natural Language Processing] methods, it might be that there is no human-understandable explanation as to why beyond: 'his corpus of discourse resembles those of others whom we thought interesting'; or the conceptual opposite: 'his discourse looks or sounds different from most people's.' " Phillip Rogaway, Professor of Computer Science, University of California (Froomkin 2015)

As Americans, we value our privacy. Yet we also value our convenience and our safety, sometimes at the peril of our privacy. We may not have the problem of millions of AI-equipped CCTV cameras on every street corner, but we've submitted to surveillance by owning one of the most useful and ubiquitous machines in history: the smart phone.

When Snowden released his trove of NSA documents to journalists, we got a glimpse behind the curtain of our government's most advanced surveillance organization, the NSA. The documents showed how they were able to listen in even on our most private conversations that we believed to be secure by installing backdoors. A decade ago, we could have written off such an endeavor as information overload; *how could they possibly sift through all that data to find anything interesting, let alone single me out?*

The answer came from machine learning and big data technologies. Apache Hadoop is capable of storing massive amounts of files and the Hadoop ecosystem provides dozens of open source tools to perform data mining on giant data sets in parallel, speeding up analyses that would have taken weeks into something that can run over night and have results ready by the time you get into work in the morning. The threat of mass surveillance has gone from tinfoil hat conspiracy theories to standard practice that we give implicit consent to by using an Internet-connected device or making a phone call.

The Snowden dump was controversial, as many warned that our intelligence methods would be exposed to terrorists and they would then change their tactics. But the cost of safety and the cost of privacy are not mutually exclusive.

> "The big question is this: how do we design systems that make use of our data collectively to benefit society as a whole, while at the same time protecting people individually? Or, to use a term from game theory, how do we find a 'Nash equilibrium' for data collection: a balance that creates an optimal outcome, even while forgoing optimization of any single facet?
>
> "This is it: this is the fundamental issue of the information age. We can solve it, but it will require careful thinking about the specific issues and moral analysis of how the different solutions affect our core values." *Data and Goliath* (Schneier 2016)

In other words, ethics.

The Need for Ethics

Through some of the previous case studies, one can begin to understand that there is a need for ethics in some form. Unfortunately, this is where it gets tricky. We begin to part ways with the binary ones and zeros of machines and stumble into the nebulous universe of philosophy. Aristotle argued in the first book of his Nicomachean Ethics that we can all agree that the goal of our actions is to attain some "good". But our agreement on what is "good" is where we differ. (Ross 2009)

Often we find disparities in our definition of "good" when it comes to our personal interests. Naturally, a corporate CEO will be in favor of reduced corporate taxes. Paying less in corporate taxes is a means to achieve an end (more profit for the company) in which the CEO benefits (via compensation). The reduction in taxes for the corporation however, must be passed on to someone, and often that ends up being lower and middle class workers, many of which may work for the company. Those workers are not likely to see the reduction in corporate taxes as "good" in the same way as the CEO. The only way to reconcile this is to get everybody on the same page of what should be considered "good".

Substituting automation in for the tax cuts in the prior example, and the story is the same. Human resources are expensive. They require an hourly wage or salary, health and retirement benefits, they require sleep, lunch and restroom breaks, and their output is inconsistent and variable. Robots on the other hand, require only electrical power and occasional maintenance.

Given the choice between a human that assembles a car imperfectly for eight hours a day and a machine that does it nearly perfectly for 24 hours a day, the manufacturer is going to go with the robots almost every time. Even if the machines are very expensive, given enough time and sales, the return on investment is bound to turn positive.

This outcome is undoubtedly good for the company and the executives, but the workers who lose their jobs will have a hard time finding any good in it. Yet, the wheel of change rolls on, and we are powerless to stop it. As painful as it is to accept this rapid change, it makes little sense to cease technological discovery. People were being displaced from their jobs during the Industrial Revolution, but if we had ceased technological advancement, our life expectancy would presumably be around 40 years, if we extrapolate out, instead of the nearly 80 years we can expect today. (Hodges 2009)

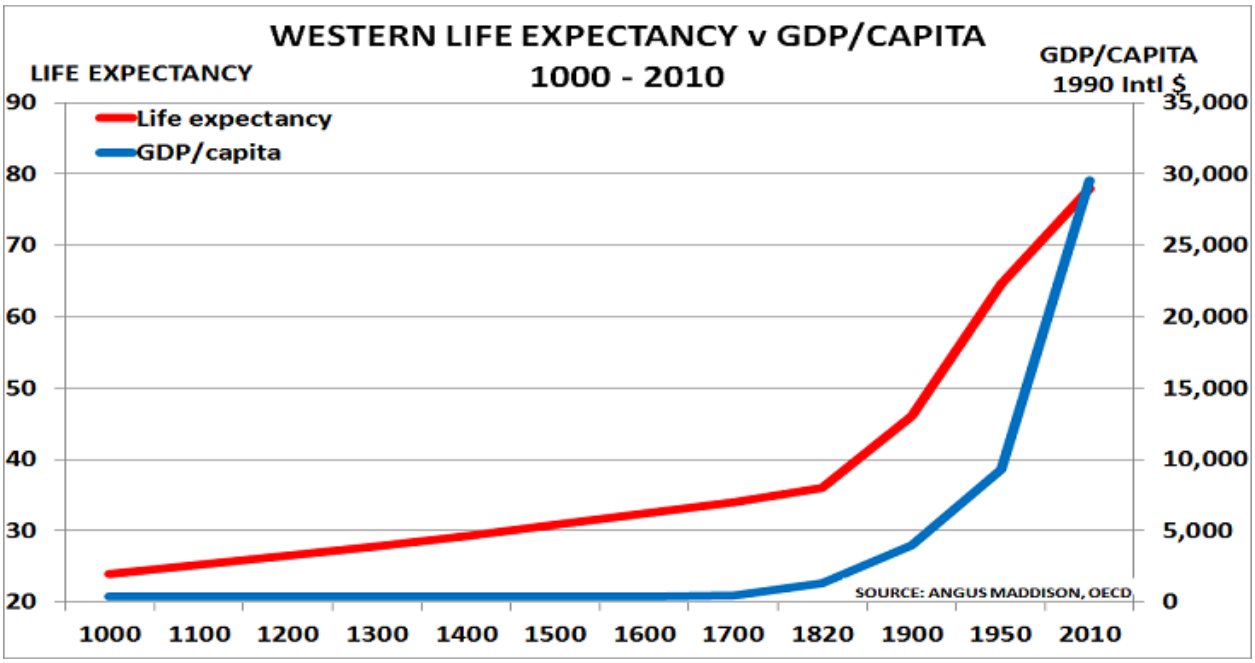

Figure 1: Human life expectancy since 1000 C.E.

Furthermore, the debate over whether technological advancement is good or bad is nonsensical. Technological evolution is itself, neutral; a product of our human achievement. How we as humans conduct ourselves with respect to the applications and advancements of technology are the details, in which, lies the Devil.

Through the previous case studies, we have seen the challenges that come with machines that are given the ability to learn and make decisions. Even in the absolute best case scenarios, where the creators of the AI have perfectly benevolent intentions, problems arise from ignorance or short-shortsightedness. Who is allowed to train the model? What is in our dataset? Is the problem we are trying to solve with machine learning legal and ethical in the first place? Who may potentially be affected by the machine's decision making, and is there potential for harm?

All of these questions can and should be answered within an ethical framework. But as we have seen time and again, ethics is usually not enough to dissuade people from causing harm to others. This is why we have laws. A speed limit is not an ethical suggestion, but rather a legal one, punishable by fines or even jail time. Of course people still break the law, but at least some of those people are caught. This serves two purposes. First, it allows us to punish those who have been caught breaking the law. And second, it serves as

a deterrent for those who may not follow ethical standards, but who are afraid of being caught and punished for breaking the law.

Regulating AI is Hard

Governments will have a hard time regulating AI. Even with Elon Musk waving the red flag, and calling for governments to start regulating AI (Vincent 2017), putting regulations around something so highly technical, and which most politicians barely have a cursory understanding of, is bound to be an uphill battle for everyone involved.

Very few congress members have the technical background needed to speak intelligently on the nuances of machine learning. In recent years, we have seen misinformed lawmakers weigh in on the perhaps equally technical matter of encryption, proposing backdoors on smart phones. But in calling for a backdoor on an encrypted device, they show how little they understand the technology, as a backdoor for one is potentially a backdoor for all. (Mossberg 2015)

There is a "Goldilocks principle" that should be pursued when regulating AI. Assuming lawmakers were successful in passing any kind of regulation, they now run into the danger of over-regulating. In October of 2012, an Italian court found six scientists guilty of manslaughter and sentenced them to six years in prison for failing to predict an earthquake that killed more than 300 people. (Sisto 2012)

Scientists argue that this sets a dangerous precedent that may very well discourage other scientists from important work for fear of retaliation for wrong answers. The Italian scientists were later exonerated (Cartlidge 2015), but such a precedent could spook many scientists and researchers from exploring potentially life saving solutions to many of the problems that plague us today.

IBM Watson, for instance, is using machine learning to try to help identify melanoma in its early stages. Such an achievement could lead to at-home fully automated checkups, and free up many dermatologists from routine skin checks, freeing up doctors and saving lives. (Codella 2016)

If we threaten to sue or jail those researchers for getting it wrong, interest in finding a solution is bound to drop rapidly. It then becomes an economics problem. Is the incentive to find technological breakthroughs higher than the potential jail time of getting it wrong? Remember, nature is stochastic. Randomness is inherent and nothing is guaranteed.

Imagine you are developing an algorithm that was a potential cancer detector. It achieves an impressive 2% Type II error rate, meaning out of every 100 people that are cancer-free, it misses two that have cancer. This would be a very useful tool, yet if you knew you could go to jail for telling the two people they don't have cancer when they actually do, you would probably not release the algorithm and more people will continue to die of cancer.

Some of the more practical cases may be covered by existing laws. HIPPA, for example, protects patients from being personally identified with any of their medical history, but does not prohibit the cross-institution sharing of anonymized health data, as is being used in the Cancer Moonshot project initiated by former Vice President Joe Biden. By following HIPPA regulations, the National Cancer Institute, Amazon Web Services, and Microsoft are able to collaborate on important research that could lead to breakthroughs in cancer research, while maintaining an ethical framework codified by law that protects individual patients' privacy. (Office of the Vice President 2016)

When it comes to government surveillance or autonomous weapons, the challenges increase. Fear is a compelling motivator for most people, and as long as they are told there is a threat and they must surrender their privacy to be kept safe, many will happily do so. What can we do about it then?

World renowned security expert, Bruce Schneier has some recommendations. In his New York Times Bestseller *Data and Goliath*, he puts forth three specific recommendations. First, make an effort to see surveillance. We can't talk about a problem if we don't notice it. Second, we should talk about it. Have open conversations with family and friends. Post about it on social media. Make your opinions known. The more conversation

there is around it, the more the issue becomes visible. And finally, organize politically. We should demonstrate that these problems are common to all of us, and the solutions are non-partisan. (Schneier 2016)

Above all, the pioneers in the commercial sector, who are forging the bleeding edge of AI need to take it upon themselves to institute strong ethical standards in their organizations. A company can enact changes to their internal policies much faster than a government can enact legislation. Integrating ethics standards into performance reviews, for instance, and creating an overall culture of high ethical standards sets a solid foundation on which AI can thrive without causing harm, be it intentional or unintentional.

Not a Solution, but a Start

The goal of this paper is not to engage in philosophical debates on ethics - what is and what isn't ethical is a very hard question to answer. (Some things are fairly universally agreed upon to be in one bucket or the other, but often, moral conundrums occupy some space on a sliding scale.)

Rather, my goal here is to argue that we need to begin thinking about it now. With the advancements in processor technology, the sophistication of machine learning algorithms has improved drastically in recent years. While it feels like we have had AI around all along, we are, in reality, at the forefront of truly useful AI.

Just ten years ago in 2007, the game of Checkers was solved by researchers at the University of Alberta. (Mullins 2007) Now, ten years later, Google defeated world class Go champion Ke Jie, 60 - 0. (Shead 2017) Go is considered one of the most difficult games for computers to win because of its huge branching factor.

As the technology advances at breakneck speeds, we must call for ethics now. There is no more time for delay. With each new technological achievement that is unlocked, ethics or the lack thereof, will have a lasting effect on how the technology is applied. Just because we *can* do something doesn't mean we *should*. When such a moral situation arises, how will we know which fork in the road to take? Without ethics, we don't.

So what do we do? How do we start? A logical starting point is where the knowledge is instilled in the first place: higher education. We should not be taught science without a set of ethics to guide us.

Most Computer Science and Management of Information Sciences programs require a course in IT Ethics. Academic institutions recognized the need to train technology students in ethics using case studies as we have done here, to describe pitfalls that students may encounter throughout their careers and how to navigate them.

Artificial Intelligence has traditionally fallen under Computer Science departments, but increasingly, this profession has an expansive set of titles and practitioners come from a wide array of backgrounds. In addition to traditional Computer Scientists, disciplines as diverse as Mathematics and Statistics, to Biology, to Political Science may find themselves in a role practicing Data Science, building machine learning algorithms in an academic setting, a small business or enterprise setting, or even government.

If we package ethics alongside statistics and machine learning classes, we can cover a broad range of disciplines with a single ethical framework that can guide students and professionals throughout a variety of scenarios.

Another suggestion is to modify reward structures. Enterprises and small businesses can instill ethics in their employees by making it a part of company culture. Adding ethical decision making as an objective on each employee's performance review shows that not only is it expected behavior to perform ethically, but it is "just how we do business around here." It normalizes it so that employees who want to behave unethically, must do so willfully and at their peril.

In fact, the entire academic industry is in need of change. One of the biggest problems plaguing scientific journals today is *p-hacking*, the frowned upon yet still widely practiced research method of finding a hypothesis that fits the desired p-value (usually $p < 0.05$).

The academic reward structure has traditionally been such that researchers were rewarded by the number of papers they publish, rather than the quality of those papers. This is changing, but peer review is still an issue (either it doesn't happen or it is done haphazardly), and researchers aren't necessarily publishing important true findings so much as a finding that hits the p-value threshold. This practice has led to a

distrust of science by the public. A large part of it can be attributed to the incentives that scientists have to publish more papers.

Incentives play a key role in ethics. Man cannot be trusted to do the right thing no matter what. Man will usually do the right thing as long as the incentive is there. So let's incentivize ethics. Reward people for doing the right thing consistently, and particularly when there was a clear choice of doing the wrong thing.

These are just a few suggestions, not an exhaustive list. Above all, we need to be thinking about ethics, and the role it plays in AI both today, and in the future. Nothing we have created as a civilization compares to the incredible potential of AI to make our lives better, worse, or some combination of both. How we want to materialize the results of our work will depend on how we go about the work in the first place.

> "If, then, there is some end of the things we do... clearly this must be the good and the chief good. Will not the knowledge of it, then, have a great influence on life? Shall we not, like archers who have a mark to aim at, be more likely to hit upon what is right? If so, we must try, in outline at least, to determine what it is, and of which of the sciences or capacities it is the object." *Nicomachean Ethics* (Ross 2009)

References

"Artificial Intelligence Will Replace Half of All Jobs in the Next Decade, Says Widely Followed Technologist." 2017. *CNBC*, April. https://www.cnbc.com/2017/04/27/kai-fu-lee-robots-will-replace-half-of-all-jobs.html.

Asimov, Isaac. 1950. *I, Robot*. Doubleday.

Berger, Eric. 2016. "Federal Report: AI Could Threaten up to 47 Percent of Jobs in Two Decades." *Ars Technica*, December. https://arstechnica.com/information-technology/2016/12/federal-report-ai-could-threaten-up-to-47-percent-of-jobs-in-two-decades/.

Cartlidge, Edwin. 2015. "Why Italian Earthquake Scientists Were Exonerated." *Science*, February. http://www.sciencemag.org/news/2015/02/why-italian-earthquake-scientists-were-exonerated.

Cellan-Jones, Rory. 2016. "Stephen Hawking - Will Ai Kill or Save Humankind?" *BBC News*, October. http://www.bbc.com/news/technology-37713629.

Chiel, Ethan. 2016. "Who Turned Microsoft's Chatbot Racist? Surprise, It Was 4chan and 8chan." *Splinter*, March. http://splinternews.com/who-turned-microsofts-chatbot-racist-surprise-it-was-1793855848.

Codella, Noel. 2016. "Identifying Skin Cancer with Computer Vision." *IBM Research*. https://www.ibm.com/blogs/research/2016/11/identifying-skin-cancer-computer-vision/.

Dodgson, Lindsay. 2017. "9 'Future-Proof' Careers, According to the World's Largest Job Site." *Business Insider*, May. http://www.businessinsider.com/careers-that-are-safe-from-automation-2017-5/#gig-worker-variable-9.

Froomkin, Dan. 2015. "The Computers Are Listening: How the Nsa Converts Spoken Words into Searchable Text." *The Intercept*, May. https://theintercept.com/2015/05/05/nsa-speech-recognition-snowden-searchable-text/.

Gilbert, David. 2017. "Russian Weapons Maker Kalashnikov Developing Killer Ai Robots." *Vice News*, July. https://news.vice.com/story/russian-weapons-maker-kalashnikov-developing-killer-ai-robots.

Grush, Loren. 2015. "Google Engineer Apologizes After Photos App Tags Two Black People as Gorillas." *The Verge*, July. https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas.

Hodges, Paul. 2009. "Rising Life Expectancy Enabled Industrial Revolution to Occur." *Chemicals & The Economy With Paul Hodges*. http://www.icis.com/blogs/chemicals-and-the-economy/2015/02/

rising-life-expectancy-enabled-industrial-revolution-to-occur/; ICIS.

Khrennikov, Ilya. 2017. "Moscow Deploys Facial Recognition to Spy on Citizens in Streets." *Bloomberg*. https://www.bloomberg.com/news/articles/2017-09-28/moscow-deploys-facial-recognition-to-spy-on-citizens-in-streets.

Lo, Tiffany. 2017. "Big Brother Is Watching You! China Installs 'the World's Most Advanced Video Surveillance System' with over 20 Million Ai-Equipped Street Cameras." *Daily Mail*, September. http://www.dailymail.co.uk/news/article-4918342/China-installs-20-million-AI-equipped-street-cameras.html.

Miller, Claire Cain. 2017. "A Darker Theme in Obama's Farewell: Automation Can Divide Us." *New York Times*, January. https://www.nytimes.com/2017/01/12/upshot/in-obamas-farewell-a-warning-on-automations-perils.html?mcubz=3&_r=0.

Mossberg, Walt. 2015. "Mossberg: An Encryption 'Backdoor' Is a Bad Idea." *Recode*. https://www.recode.net/2015/12/2/11621080/mossberg-an-encryption-backdoor-is-a-bad-idea.

Mullins, Justin. 2007. "Checkers 'Solved' After Years of Number Crunching." *New Scientist*. https://www.newscientist.com/article/dn12296-checkers-solved-after-years-of-number-crunching/.

Nesbit, Jeff. 2015. "We All May Be Dead in 2050." *U.S. News and World Report*, October. https://www.usnews.com/news/blogs/at-the-edge/2015/10/29/artificial-intelligence-may-kill-us-all-in-30-years.

Office of the Vice President. 2016. "FACT Sheet: Vice President Biden Delivers Cancer Moonshot Report, Announces Public and Private Sector Actions to Advance Cancer Moonshot Goals." https://obamawhitehouse.archives.gov/the-press-office/2016/10/17/fact-sheet-vice-president-biden-delivers-cancer-moonshot-report.

Pomerleau, Mark. 2015. "Next Phase of Blue Force Tracking Hits the Ground Running." *Defense Systems*, March. https://defensesystems.com/articles/2015/03/05/army-blue-force-tracking-jbc-p.aspx.

Reuters. 2017. "Man Killed in Tesla 'Autopilot' Crash Got Numerous Warnings: Report." *CNBC*, June. https://www.cnbc.com/2017/06/20/man-killed-in-tesla-autopilot-crash-got-numerous-warnings-report.html.

Rodriguez, Ashley. 2016. "Microsoft's AI Millennial Chatbot Became a Racist Jerk After Less Than a Day on Twitter." https://qz.com/646825/microsofts-ai-millennial-chatbot-became-a-racist-jerk-after-less-than-a-day-on-twitter/.

Ross, W.D. 2009. *Aristotle: Nicomachean Ethics (Translated by W.D. Ross)*. Vol. 1. http://classics.mit.edu//Aristotle/nicomachaen.1.i.html; The Internet Classics Archive.

Sasse, Ben. 2016. "Trade, Automation, and the Transformation of Work." *Twitter*. https://twitter.com/SenSasse/status/806889568276336641.

Schneier, Bruce. 2016. *Data and Goliath*. W. W. Norton & Company.

Shead, Sam. 2017. "DeepMind Is on the 'Charm Offensive' for Google in China." *Business Insider*. http://www.businessinsider.com/google-deepminds-alphago-ai-beat-the-best-go-player-in-the-world-in-its-first-game-2017-5.

Sisto, Alberto. 2012. "Italian Scientists Convicted over Earthquake Warning." *Reuters*, October. https://www.reuters.com/article/us-italy-earthquake-court/italian-scientists-convicted-over-earthquake-warning-idUSBRE89L13V20121022

Staff, RBR. 2012. "Robot Freedom: Electromobility from High-Energy Batteries." *Robotics Business Review*, May. https://www.roboticsbusinessreview.com/uncategorized/robot_freedom_electromobility_from_high_energy_batteries2/.

Stecher, Benjamin. 2017. "In the World of Tomorrow, Google Plans to Use Ai to Do Everything." *Futurism*, April. http://www.bbc.com/news/magazine-40124781.

Tangenes, Gisle R. 2004. "Peter Wessel Zapffe: The Last Messiah." *Philosophy Now* 45 (March). https://philosophynow.org/issues/45/The_Last_Messiah.

"Testing Tesla's Autopilot System at 70mph." 2015. *Car Throttle*. https://www.youtube.com/watch?v=tP7VdxVY6UQ.

Various. 2015. "Autonomous Weapons: An Open Letter from Ai & Robotics Researchers." *Future of Life*

Institute, July. https://futureoflife.org/open-letter-autonomous-weapons/.

Vincent, James. 2017. "Elon Musk Says We Need to Regulate AI Before It Becomes a Danger to Humanity." *The Verge*, July. https://www.theverge.com/2017/7/17/15980954/elon-musk-ai-regulation-existential-threat.

West, John. 2016. "Microsoft's Disastrous Tay Experiment Shows the Hidden Dangers of AI." *Quartz*, April. https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/.

n.d. *Association of American Railroads*. https://www.aar.org/policy/positive-train-control.