**Week 6: Ethical Challenges with AI**

Nate Bachmeier

April 14th, 2019

TIM-8101: Principals of Computer Science

Northcentral University

# Ethical Artificial Intelligence Design

What does it mean for systems to be designed without intentional ethical frameworks? Perhaps this conjures thoughts of Terminator’s *Skynet* as we battle the machines wishing they followed Isaac Asimov’s *Three Laws of Robotics*. In a way those are easier battles to fight as the difference between right and wrong is very clear. In today’s world designing ethical systems lives more in the grey zone, where we “know it when we see it” but cannot always define it.

Consider a facial recognition system which accounts for demographic information, does that make it immoral? Perhaps that information is used to choose between specialized models for persons of different ancestries. The specialized model could be used to overcome certain facial blindness scenarios, improving reliability and end user satisfaction.

## Many Worlds, Many Rules

What if the demographic information is used by a college enrollment department? This becomes even more complex as different societies have different perspectives. European’s nations would raise litigation against the product team. In American societies they would call this affirmative action, a necessary tool to equalize the playing field.

Onufn describes this challenge as the Many World’s Many Rules problem of morality. He states that “if one is to build a system that models ethical behavior then it is critical to understand different implementations of the model (Onuf, 2016).” This means that rich contextual information is required as a prerequisite.

There are three worlds our world; the rest of the world; and the largest world possible. Each of these different societal scopes will have different bylaws that must be adhered to. For example, a person’s medical records can be shown to the end user in full detail. If that same information changes scope, then it might only be ethical to show them in aggregate form.

## One World, One Set of Rules

Recently the Europe took a different position on the broader topic of ethical AI, stating that there is only one world. In that world systems need to strictly adhere to the notions of privacy, transparency, safety and accessibility (European Commission, 2019).

As the second phase of the Generate Data Protection Regulation (GDPR), their end goal is to provide a guiding hand in system design. This comes after multiple high-profile scandals, such as Facebook and Cambridge Analytics, have demonstrated an inability to do the right thing on their own. Experts agree “Europe [is] at the forefront of the regulation and reflections on these issues (Inversardi, 2019).”

## The Fallacy of Privacy

When Joseph Sirosh became vice president of Microsoft’s AI division a Q/A session was held. One of the questions that really resonated with me as an observer was “how do we address the privacy concerns caused by AI?”

He replied that the traditional idea of privacy is dead, and the conversation needs to transition toward responsible governance of information. Adding that humans inherently lack enough entropy for many common scenarios.

Consider a typical day, a person will get up in the morning and drives to work. After roughly 8 hours they commute to a restaurant or grocery store and then retire the night at home. If we provide this model to a machine and then give it a stream of GPS coordinate updates, the system can accurately predict many aspects of that person’s life. Not only simple things, such as food recommendations but deeply personal dimensions such as are they cheating on their spouse.

## Doing Good vs Avoiding Bad

Psychologies ask the question why should actors be ethical in the first place? Though empirical studies they suggest that ethical choices can be modeled on a 2x2 matrix of: Give to do good or avoid bad; and Equalize to do good or avoid bad (Tappin & Capraro, 2018).

For example, if an autonomous medical system euthanizes their patient can it still be considered ethical? The question raises memories of debates over Dr. Jack Kevorkian, and his assisted suicides. An argument can be made that the system without this feature is more ethical, as it “avoids bad.” Similarly, an argument can be made that it needs the feature to be ethical to “give good.”

Then consider if the medical system has a larger holistic view and decides that an elderly patient needs to die, so that their organs can be provided to a younger patient. By equalizing for good it must be ethical. However, even doing right can be wrong unless those decisions can be audited and are fully transparent to an oversight committee (Matsuzaki, 2017).

## Giving Medical Information to Machines

One of the challenges for professionals is keeping track of the ocean of biological information that is needed to medical procedures.

First diagnosticians needed to come up with an explanation of the illness. Perhaps they reviewed medical imaging results to come to that conclusion. Yet studies shown that the same image given the same doctor multiple times result in different conclusions (Thaler, 2008). However, an artificial intelligence system will give the same answer to the same image each time it is asked.

Additionally, those doctors need to consider different demographic information as certain groups of people are more resistant to different treatments. Finding the intersect between ever changing research results and generic markers quickly becomes a big data problem, and more suited for machines.

Finally, the patient is ready for surgery and a team of nurses monitors a small number of sensors. This approach will not scale into the 5th generation of wireless where vast sensor networks become readily available. Instead artificial intelligence will monitor thousands of data points in real time and detect potential issues earlier.

## Challenges in Reliability of Medical AI

Which would a patient choose between an autonomous surgeon, with an impressive 1% mortality rate, versus an expert human with a 92% success rate? Does the calculus change if they know the doctor will be held liable for killing them, versus the machine would not shed a tear?

Current legislation is not keeping up with technological advancements and there is limited legal recourse for a mistake. Like the gun industry the manufacture is not responsible provided they give advanced warning (Kasprak, 2018) (Matsuzaki, 2017).

## Challenges in Privacy of Medical AI

To train artificial intelligences models requires many patient history records. Do the patients have a right to privacy even if the data is anonymized? This is not a new question within the medical community as can be seen in the case of Henrietta Lacks. With limited consensual understanding a doctor collected samples of Lack’s cervical cancer cells (Skloot, 2010). Today HeLa cells are the basis of modern cancer research, have generated billions of dollars for pharmaceutical companies, and saved millions of lives. Clearly several positive outcomes arose from that decision, yet none of them were to the benefit of the patient.

## Challenges in Security of Medical AI

Multi-Agent Systems (MAS) perform an ethical judgement process by performing an intersect of desires, beliefs, and possible actions. The resulting set is then scored based on morality rules, rightness, and finally contextual judgement (Cointe, Bonnet, & Boissier, 2016). For example, immoral to steal but permissible to steal bread to feed a hungry child.

As the systems can be make decisions are with larger contextual information it will push the systems into these grey zones. Consider a surgical system that is performing an operation on a genocidal dictator. If the system kills the patient and manipulates the audit log, then it will save countless lives. Based on the scoring algorithm that could be a very logical decision. In today’s world the human doctor would be thrown in jail, that is hardly a punishment for the machine.

The machines turning against us is not the only security concern, as malicious hackers could compromise the equipment to either kill the patient or steal the medical information. According to Wired, these vulnerabilities exist in today’s systems and need to be fixed before the humans are removed completely (Zetter, 2014) (Newman, 2017).

Instead of wearing the tin foiled hat the systems need to be designed to mitigate these scenarios. Using distributed ledgers such as the ones used by cryptocurrency, the audit logs can be designed cryptographically verifiable.

Vulnerabilities in hardware and software systems will always exist and needs to be combatted through more artificial intelligence systems, not less. Humans are unlikely to notice the slow exfiltration of a medical database, yet AI systems can be built to detect and respond to the anomaly.

# Where is all this headed?

# Determine controversial area of research

Artificial intelligence and autonomous systems are defined as any component that can derive contextual meaning without human interaction (Inversardi, 2019). For many this conjures images of *Skynet* from the Terminator series, where our only hope is an implementation of Isaac Asimov’s Three Laws of Robotics.

As these systems do not require humans there is a natural fear that they will supersede their creators. The fear of losing control leads to a desire to stifle that creative force and limit the areas that are directly impacted.

One controversial area where this can be seen is the health care industry (Kostopoulos, 2016) (Matsuzaki, 2017). Health care will eventually become enriched through the adoption of AI systems however it must first overcome some key challenges in reliability, transparency, liability, security, and privacy.

# Why is it controversial

Imagine that a loved one needs to have a complicated surgery, and the doctor tells their patient that there are two choices. The first is an excellent surgeon with an 8.9% mortality rate, or a machine that commands a 1% mortality rate.

If something goes wrong the machine will not adapt to the changed conditions, nor can the patience’s family sue when things go wrong. When the patient dies who should be held accountable-- the manufacture or the hospital? Modern legal systems are not keeping pace with technological advancements which introduces challenges. Recent court cases have found that gun manufactures are not liable for shootings (Kasprak, 2018), can the same be said for autonomous devices?

When the machine decides to perform an action during the surgery, how can we know that it is the right one to make? What systems are auditing those choices and what evidence can be provided to a review board to justify those behaviors?

For these machines to make intelligent decisions they require access to large amounts of data. How does that impact the patients right to privacy versus the safety across the larger community?

This is not a new challenge as can be seen with Henrietta Lacks and the theft of her cancerous cells (Skloot, 2010). Her cells enabled numerous medical discoveries and saved countless lives, however her personal privacy was forfeit.

Assume for a moment that only a subset of patient data was made available to the AI system. This would lead to more challenges as the reliability would be reduced. How would the machine react when presented with a completely foreign configuration? The human can reach out to colleges or consult medical encyclopedias, neither of these actions are freely available to the machines.

On top of these challenges are concerns related to the security of the machines. Malicious actors could manipulate the firmware in such a way as to kill the patient (Zetter, 2014). Aside from a *Manchurian Candidate* scenario, the same is not true of the human.

# What solutions are being presented

When Joseph Sirosh became vice president of Microsoft’s AI division a Q/A session was held. One of the questions was “how do we address the privacy and ethical implications of AI systems?”

He replied that the traditional idea of privacy is dead, and the conversation needs to transition toward responsible governance of information. Adding that humans inherently lack enough entropy many common scenarios. For instance, systems can determine from a stream of GPS notifications where a given end user lives, works, and privately walks their dog.

In contrast the European Commission has recently stated that personal privacy needs to be the heart of the discussion. Systems need to be built with core tenants of transparency, accountability and oversight (European Commission, 2019). They envision a world where data lives on private islands completely controlled by the end user. For instance, user’s have a right to be forgotten and delete data they have generated.

These extremes can be modeled within the *Many Worlds Theorem*, which states that there are three worlds-- our world, the rest of the world, and the largest possible world (Onuf, 2016). Onuf proposes that each societal scope needs to handle different permissions and granularity of information.

For example, within my personal view health dashboard key performance indicators (KPI) need to be highly personalized and confidential. However, with the rest of the users of the service those results should be anonymized. If that data leaves the control of the service provider, then it is given in aggregate format.

# What is the ultimate resolution to these controversies

Businesses and government

# What are future avenues for research?

One of the biggest challenges with ethical system design is that consistent modeling frameworks do not exist. There are basic human rights that most societies agree upon, but a lot of grey area exists.

Consider the scenario where many autonomous devices are operating in a shared space. It can be proven that it is more efficient for the all devices to collaborate (Rabanal & Friedman, 2015), but there does not enforce all systems to behave as such.

There is also a disconnect that these systems are not being ethical because it is the right thing to do, they are being ethical as a means of avoiding bad. They are designed to be pro-sociality.