**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).”

Not all debacles require legislative interference, as seen with Microsoft’s Tay bot, an artificial intelligence that learned human interactions through Twitter. Within 16 hours the system became an angry troll and Holocaust denier (Reese, 2016). An apology was made, and the system decommissioned. An interest point of that story which wasn’t wildly reported is that a version of Tay had been deployed in Asian social media for over a year. This also serves as a prime example of how societal norms have a huge influence over AI algorithms.

## 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 becoming sentient and turning against us is not the only security concern. 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 a tin foiled hat, autonomous systems need to be designed to mitigate such scenarios. Using distributed ledgers such as the ones used by cryptocurrency, the audit logs can be designed cryptographically verifiable. Firmware on these devices needs to also be signed to prevent tampering.

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?

At the beginning of the Industrial Revolution there was a general fear that electricity would be the end of humanity. Now that we are entering the Super Intelligence Revolution, there are similar questions as humans fear becoming obsolete or slaves to robotic masters (WhyFuture, 2017).

## Path to Adoption

Along the path to adoption road blocks will be created, in a misguided effort to protect jobs or ensure the privacy of citizens. Eventually these fears will subside, and humans will see these autonomous machines for what they really are, merely another tool in their toolbox. The potential benefits far outweigh the risks of not adopting these technologies.

That is not meant to discount the impact of unethical system design. These systems lack emotions and an inherent sense of right versus wrong. They only know how to optimize highly dimensional data into a model. If society expects that to occur in an ethical manner, then the corporations that build these systems need to be held accountable for their creations.

## Future Research Topics

It seems that every step forward raises addition challenges and identifies new research topics that need to be addressed.

First, how will we hold these businesses accountable? Despite the missteps of Facebook both their user base and sales revenues are the largest yet. If we are overly restrictive then scientific discovery is impeded.

Second, how do we protect the rights of an individual without impeding scientific discovery? Europe has gone to great lengths to implement the right to be forgotten, but as the data becomes anonymized and transformed it really removed?

Third, can a generalized framework for ethical design be exposed as part of the core operating system? Many efforts have created qualitative models for making ethical judgements decisions, but a need for quantitively reasoning also exists.

Fourth, how do we design ethical systems which can adapt variable levels of restrictive societal context? Limiting to the lowest common denominator will slow down advancements, yet the system needs to be respectful of its user base.

# References

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical Judgment of Agents’ Behaviors in Multi-Agent System. *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems* .

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical Judgment of Agents’ Behaviors in Multi-Agent Systems. *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems* .

European Commission. (2019, April 8). *Ethics guidelines for trustworthy AI.* Retrieved from European Commission: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Inversardi, P. (2019). The European Perspective on Responsible Computin. *Communications of the ACM Vol. 62*, 62-69.

Kasprak, A. (2018, February 21). *Gun Manufacturers Can’t Be Sued For Crimes Committed With Their Products?* Retrieved from Snopes: https://www.snopes.com/fact-check/gun-manufacturers-crimes-products/

Kostopoulos, L. (2016). The Emerging Artificial Intelligence Wellness. *California Wester Law Review Volume 55*, 235-254.

Matsuzaki, T. (2017). Ethical Issues of Artifical Intelligence in Medicine. *California Western Law Review Volume 55*, 255-274.

Newman, L. (2017, March 2). *Medical Devices Are the Next Security Nightmare*. Retrieved from Wired: https://www.wired.com/2017/03/medical-devices-next-security-nightmare/

Oana, O., Cosmin, T., & Valentin, N. (2017). Artificial Intelligence - A New Field of Computer Science. *Economic Sciences Series Volume XVII, Issue 1*, 356-360.

Onuf, N. (2016). Many Worlds, Many Theories, Many Rules: Formulating an Ethical System for the World to Come. *Revista Brasieira de Politica Internacional*.

Rabanal, J., & Friedman, D. (2015). How Moral Codes Evolve in a Trust Game. *MDPI Games*, 150-160.

Reese, H. (2016, March 24). *Why Microsoft's 'Tay' AI bot went wrong.* Retrieved from Tech Republic: https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

Skloot, R. (2010). *The Immortal Life of Henrietta Lacks.* Crown.

Tappin, B., & Capraro, V. (2018). Doing good vs. avoiding bad in prosocial choice: A reﬁned test and extension of the morality preference hypothesis. *Journal of Experimental Social Psychology*, 64-70.

Thaler, R. (2008). *Misbehaving.* Nudge.

WhyFuture. (2017, February 9). *The Future of Artificial Intelligence and Ethics on the Road to Superintelligence.* Retrieved from YouTube: https://www.youtube.com/watch?v=8Ja\_7Fx2MmU

Zetter, K. (2014, April 25). *It's Insanely Easy to Hack Hospital Equipment.* Retrieved from Wired: https://www.wired.com/2014/04/hospital-equipment-vulnerable/