**Week 6: Ethical Challenges with AI**

Nate Bachmeier

April 14th, 2019

TIM-8101: Principals of Computer Science

Northcentral University

# Ethical Artificial Intelligence Design

What would it mean for artificial intelligence systems to be designed without ethical considerations? Perhaps this conjures up 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 ethical systems live more in the grey zone, where we “know it when we see it” but cannot always define it.

If a facial recognition system uses demographic information, does that make it immoral? Perhaps that information is used to choose between specialized models for persons of different ancestries. The specialized models could improve reliability and end user satisfaction by overcoming certain facial blindness scenarios.

## 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 enrollment team. In American societies they might see call this as affirmative action-- a necessary tool for equalizing the playing field.

Onufn describes these contextually specific scenarios as the *Many World’s, Many Rules* challenge 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).” These different implementations of societal norms form the basis that all morality decisions must expand.

The model proposes three worlds: our world; the rest of the world; and the largest world possible. Each of these worlds, or societal scopes, will have different bylaws that must be adhered to. For example, a person’s medical records can be shown to them in full detail. If that same information in presented in another world, then it might be ethical to only show in aggregate form.

## One World, One Set of Rules

Recently 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 when Facebook and Cambridge Analytics have demonstrated an inability to do the right thing on their own. Experts tend to agree “Europe [is] at the forefront of the regulation and reflections on these issues (Inversardi, 2019).”

Not all debacles have required 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 to that story is that a version of Tay had been deployed to an Asian social media service for over a year. This also serves as an example of how societal norms have a huge influence over AI algorithmic results.

## 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 asked 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 gets up in the morning and drives to work. After roughly 8 hours they commute to a restaurant or grocery store and then retire for 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 if they are cheating on their spouse.

## Doing Good vs Avoiding Bad

Psychologist ask the question why actors are 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. Though at the same time an argument can be made that the feature is needed to ethically 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).  
 These scenarios raise an interesting research question, how can we express right from wrong to machines? Often an example toward ethical behavior can also be a counter example. This connects back to the need for contextual information and an understanding of societal norms.

## Giving Medical Information to Machines

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

First diagnosticians needed to come up with an explanation of an illness. Perhaps they reviewed medical imaging results to come to that conclusion. Yet studies have shown that the same image given the same doctor multiple times may result in different conclusions (Thaler, 2008).

Additionally, those doctors need to consider different demographic information as certain groups of people are more resistant to various treatments. Finding the intersection between ever changing research results and genetic 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 need to monitor thousands of data points in real time to detect potential issues earlier.

## Challenges in Reliability of Medical AI

Which would a patient prefer (1) an autonomous surgeon, with an impressive 1% mortality rate, or (2) an expert human with a 92% success rate? Does the calculation change if they know the doctor can be held liable for killing them, while 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, manufactures of autonomous medical devices are not responsible for mortalities provided they gave warning in advance (Kasprak, 2018) (Matsuzaki, 2017).

To train artificial intelligences models, requires numerous copies of patient history records. Without enough training data, the results will be overfitted and not reliable. There are challenges for researchers to gain access to enough data, especially for certain rare diseases.

## Challenges in Privacy of Medical AI

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, generating billions in revenue 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.

Graepel, Lauter, and Naehrig proposed a method for performing machine learning on encrypted values. They argue that AI does not need the real values until the final steps of the process. Their solution is significantly more computationally complex but does protect the privacy for the user from end to end (Graepel, Lauter, & Naehrig, 2012).

## 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, it is immoral to steal but socially acceptable to steal bread to feed a hungry child.

As systems being asked to make decisions with more contextual information it will lead them 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 MAS scoring system that could be a very logical decision. In today’s world the human doctor would be thrown in jail, which is hardly a punishment for the machine.

Machines becoming sentient and turning against us is not the only version of that scenario. 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 human oversight can be removed (Zetter, 2014) (Newman, 2017).

Instead of wearing a tin foiled hat, we need to design autonomous systems with mitigations toward such scenarios. Both the audit logs and firmware could be cryptographically signed to prevent tampering. Unlike in science-fiction novels, the machines are not going to rise against us, they are simply traversing a model and reacting to event streams.

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 such anomaly. In recent software engineering conferences, it has been repeatedly demonstrated that AI can find security vulnerabilities more efficiently than humans (Lemieux, Sen, Padhye, & Song, 2018) (Cummins, Petoumenos, Leather, & Murray, 2018).

# 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 and to ensure the privacy of citizens. Eventually these fears will subside, and humans will see these autonomous machines for what they really are-- another tool in the 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

The expansion of artificial intelligence into every area of research and human interaction will open many doors for research topics. It is important that each of them be answered through ethical design and implemented in a manner that is respectful to the end users.

First, how will we hold these businesses accountable? Despite the missteps of Facebook both their user base and sales revenues are the largest yet.

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 quantitative reasoning also exists.

Fourth, how do we design ethical systems to adapt to variable levels of restrictive societal context? The needs and requirements between different worlds, will be worlds apart.

# 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* .

Cummins, C., Petoumenos, P., Leather, H., & Murray, A. (2018). Compiler Fuzzing through Deep Learning. *ISSTA’18, July 16–21, 2018, Amsterdam, Netherlands*.

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

Graepel, T., Lauter, K., & Naehrig, M. (2012, December). *ML Confidential: Machine Learning on Encrypted Data.* Retrieved from Microsoft Research: https://www.microsoft.com/en-us/research/publication/ml-confidential-machine-learning-on-encrypted-data-2/

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.

Lemieux, C., Sen, K., Padhye, R., & Song, D. (2018). PerfFuzz: Automatically Generating Pathological Inputs. *Proceedings of 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*.

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/