Week 7: Ethics and AI

Nate Bachmeier

TIM-8150: Artificial Intelligence

November 15, 2020

Northcentral University

# Ethics and AI

Artificial intelligence is a scary black box that spreads malicious propaganda, destroys jobs, and seeks to undermind honest citizens’ values. This statement is intentionally farcical, yet it also touches on real concerns of ethical AI designs. People fear what they do not understand and use science-fiction to fill these gaps. Within those futuristic worlds, machines become the dominant species that controls every decision of an enslaved human population. However, several challenges prevent this transition of power from becoming a reality, such as intelligent systems lacks actual *intelligence* (Wildberger, 1996; Hole & Ahmad, 2019; Upchurch, 2018). Instead, organizations need to assess these tools rationally, explore applications that enhance human capabilities, and remove nondifferentiating overhead.

# Roles of Artificial Intelligence

Despite artificial intelligence already being well-entrenched in everyday life, there are concerns around its role. First, does the advancement of machine learning mean fewer jobs? Second, of those remaining jobs, are humans giving away control unnecessarily? Third, are those machines capable of manipulating the general public to steal control?

## Role in Employment

Before 1949, digging a ditch would take hours or even days with a crew of manual workers. After the invention of the backhoe, these jobs required less time with fewer employees. From the organization’s perspective, these efficiencies translate into faster time to market at lower costs. Meanwhile, the former diggers became displaced into new roles that repairs, operate, and supervises the equipment. Each of these positions comes with supply chains of supporting requirements. For instance, it takes factories to produce the backhoe parts, each staffed with hundreds of blue-collar jobs. Cities must also build universities and technical schools to train team members that will fill these roles, further expanding the job market.

Similarly, modern businesses are actively seeking methods that reduce costs and improve efficiencies through automation. The most powerful artificial intelligence applications use machines to enhance human capabilities rather than replace them (Heer, 2019; Boire, 2017). For instance, a person can write a more profound business case than a machine; however, the same machine will have fewer grammatical errors. This dichotomy exists because humans specialize in contextualizing thought versus automation uses patterns to make predictions (Schleer et al. 2019). Many professions exist as a combination of decision-making, pattern recognition, and mechanical tasks. Expert systems address specific aspects of the job requirements; however, superseding the soft-skills that unify these role components is challenging (Huang et al., 2019).

Specific low-skilled jobs, such as bank tellers and office clerical staff, are at risk of being replaced (Hamid et al., 2017). Similarly, expert pattern matching tasks like identifying tumors in MRI (Magnetic Resonance Imaging) becomes commoditized. Given the lower entry barrier, some low-skilled workers will transition to better-paying jobs that operate those sophisticated and commoditized systems. For instance, many workers cannot access foreign markets due to language and communication limitations. Artificial intelligence can aid these in these translation scenarios while leaving control with humans.

## Role in Decision Making

Many decision-making processes can benefit from machines providing recommendations and validations. For instance, a court judge could use an intelligent system to assess how their sentencing aligns with existing norms. Perhaps the machine predicts the defendant should receive five years of probation, while a judge considers fifteen years in prison. When the validation check expresses such a difference in opinions, it could suggest that unconscious bias is taking place and warrants additional considerations. That bias either provides ammunition for appeals processes or incarcerates people unjustifiably long.

While this approach has much potential, there are concerns that professionals arbitrarily accept recommendations. However, these challenges occur everywhere that automation controls the ‘last mile’ of decision making. If the suggestion comes from a machine or peer, the person in charge of the process must be accountable for the final call. Blindly delegating control to machines is dangerous, precisely because learning algorithms being greedy, brittle, rigid, and opaque (Hole & Ahmad, 2019). Until artificial brains can rationalize abstract thought, humans must perform this task.

## Role in Manipulation

Modern censorship does not restrict free speech; instead, it increases the noise and drowns the signal (Thomas, 2019). Fundamentally, marketing campaigns and propaganda machines follow the same process of Segmentation Targeting and Positioning (STP) (Kane, 2019). Delivering on this objective requires pattern matching, content delivery, and human intuition. Automation is well-suited for these tasks and can use social media channels, like Facebook and Twitter, to connect with billions of people and manage significant portions of those interactions.

Congressional and media sources raise ethical questions around the ease of access to these capabilities for political manipulation. Unfortunately, these questions are mostly talking points rather than a call for action. Artificial intelligence comes with many abstract concepts that do not fit within the complex and opaque legal language (Guiffrida et al., 2018). For instance, machines cannot reason about their instructions, so can the courts hold *it* accountable? Perhaps the system designers should be responsible for their creations. However, the algorithms are primarily algebraic formulas controlled by end-users. Without a mechanism to define and enforce a standard operating behavior, it is impossible to expect a different outcome.

# Design Considerations

Two attempts to define this mechanism are the European Union’s Ethics Guidelines for Trustworthy AI and the OECD’s Principals of Ethical AI (EU, 2019; OECD, 2019). Both documents describe the need for artificially intelligent systems to be human-centric, transparent and explainable, robust, and secure.

## Human-Centric

Robotics’s Three Law states that automation should not injure humans, ignore people’s commands, and protect their existence (Asimov, 1942). These rules lay a foundation of ideas that devices exist to cooperate and enhance humanity. Unfortunately, the machines cannot reason and therefore are slaves to their program designs. Since machines cannot devise these criteria independently, it becomes the system engineers’ responsibility to enforce these requirements. Those decisions are predominately a matter of business priorities and will vary across different use-cases. For instance, Lockheed Martin, a military weapons designer, views its human-centric role as protecting American interests at foreign nations’ expense.

## Transparent and Explainable

Artificial brains often rely on deep learning techniques through neural network solutions. These networks approximate a function that maps inputs and outputs through multiple non-parametric transforms. While data scientists can perform experiments to verify the model’s accuracy, they often cannot explain it (Gilpin et al., 2018). This limitation prevents broader adoption in places like the European Union, where the General Data Protection Regulation (GDPR) grants citizens a Right to Explanation.

Further complicating matters, neural networks learn the patterns we *ask*, not necessarily the ones we *mean*. For instance, Beauty.ai, an algorithm for rating female attractiveness, lost credibility due to only giving high scores to light-skinned candidates (Upchurch, 2018). Presumably, this outcome was not intentionally malicious, but the byproduct of not sufficiently representing minorities in the training set. Similar imbalanced issues occur across many real-world domains and require sophisticated data handling strategies (Kaur et al., 2019). Even with expert data scientists, it is possible to miss these edges cases and produce invalid predictions.

## Robust and Secure

Engineers that become data scientist follow a different curriculum than their peers that become security specialist. This distinction in training is most evident in the lack of controls across artificially intelligent solutions (Lin et al., 2018; Sethi & Kantardzic, 2018). Malicious actors can influence these predictors decisions by either inserting erroneous samples into the training set or directly attacking the probability distributions. For instance, researchers have shown that applying tiny amounts of distortion to images can cause classifiers to change classes (e.g., cat versus dog) (Sethi & Kantardzic, 2018). If people cannot trust the classification algorithms integrity, how can mission-critical environments effectively use them?

# Conclusions

Artificial intelligence is a tool that can automate mechanical tasks, pattern matching data, and enhancing human capabilities. Organizations can use these means to improve efficiency and reduce wastefulness. These innovations deprecate the need for specific skill sets and lower the entry barrier into other expert systems. While this causes an initial decrease number of jobs necessary, entirely new industries follow shortly afterward. When a society can replace low-paying jobs with high-paying alternatives, this promotion justifies the short term pain.

Machine learning technology is too immature to delegate business-critical decisions. Instead, professionals should consider these technologies for initial recommendations and to verify their choices are free of unconscious biases. For example, a court judge should assess their sentencing aligns with a regression algorithm prediction, not blindly issue that verdict. Humans must maintain control of our actions and consequences. However, it can be challenging to prevent machines from manipulating our free will.

Laws cannot keep up with technology’s high-velocity innovation, causing businesses to define and self-regulate their ethical behavior. Without an official solution for maintaining accountability, this ethical desire must compete against existing business priorities. Those priorities will vary significantly between organizations, as even defining ‘human-centric systems’ is ambiguous. Moving past those challenges are issues with the fundamental integrity of neural network technologies. Implementing transparency and explainability are open research problems outside of trivial systems. After solving those issues, ensuring the training data is inclusive requires significant investments into unverifiable results.

These limitations bring the discussion around full circle to the beginning. Artificial intelligent systems are not ethical, evil, or corruptive. They are tools that automate everyday tasks and lower the barrier to entry. Users of that tool need to be cognizant of what these predictions mean and how they influence decisions. However, that is not the same thing as delegating control with impunity. After all, in a world that lacks accountability and legal enforcement, would you trust big business to do the right thing?

# References

Asimov, I. (1942). *Runaround.*

Boire, R. (2017). Artificial Intelligence, automation, and its impact on data science. *IEEE International Conference on Big Data* (pp., 3571-3574). Boston, MA: Institute of Electrical and Electronics Engineers. doi:10.1109/BigData.2017.8258349

EU. (2019, April 8). *European Union Guidelines for Trustworthy AI*. Retrieved from Europa Futurium: https://ec.europa.eu/futurium/en/ai-alliance-consultation

Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: an overview of the interpretability of machine learning. *5th International Conference on Data Science and Advanced Analytics* (pp. 80-89). Turin, Italy: Institute of Electrical and Electronics Engineers. doi:10.1109/DSAA.2018.00018

Giuffrida, I., Lederer, F., & Vermeys, N. (2018). A legal perspective on AI trials and tribulations: how artificial intelligence, the internet of things, smart contracts, and other technologies will affect the law. *Case Western Reserve Law Review, 68*(3), 747-781. Retrieved from https://search-ebscohost-com.proxy1.ncu.edu/login.aspx?direct=true&db=tsh&AN=130372095&site=eds-live

Hamid, O., Smith, N., & Barzanji, A. (2017). Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence. *15th International Conference on Industrial Informatics* (pp. 899-904). Emden, Germany: Institute of Electrical and Electronics Engineers. doi:10.1109/INDIN.2017.8104891

Heer, J. (2019). Agency plus automation. *Proceedings of the National Academy of Sciences of the United States of America, 116*(6), 1844-1850. doi:10.1073/pnas.1807184115

Hole, H., & Ahmad, S. (2019). Biologically driven artificial intelligence. *Computer, 52*(8), 72-75. doi:10.1109/MC.2019.2917455

Huang, M., Rust, R., & Maksimovic, V. (2019). The feeling economy: managing in the next generation of artificial intelligence. *California Management Review, 61*(4), 43-65. doi:10.1177/0008125619863436

Kane, T. (2019, March). Artificial Intelligence in Politics: Establishing Ethics. *Technology and Society Magazine, 38*(1), 72-80. doi:10.1109/MTS.2019.2894474

Kaur, H., Pannu, H., & Malhi, A. (2019). A systematic review of imbalanced data challenges in machine learning: applications and solutions. *ACM Computing Surveys, 52*(4), 1-36. doi:10.1145/3343440

Lin, Z., Xiao, F., Sun, Y., Ma, Y., Xing, C., & Huang, J. (2018, April). Secure Encryption-Based Malware. *Transactions on Internet and Information Systems, 12*(4), 1799-1818. doi:10.3837/tiis.2018.04.022

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. Retrieved from OECD Principals of Ethical AI: https://www.oecd.org/legal/legal-instruments.htm

Sandybayev, A. (2018). Artificial Intelligence: Are we all going to be unemployed. *Fifth HCT Information Technology Trends* (pp. 23-27). Dubai, United Arab Emirates. doi:10.1109/CTIT.2018.8649521

Schleer, P., Drobinsky, S., de la Fuente, M., & Radermacher, K. (2019). Toward versatile cooperative surgical robotics: a review and future challenges. *International Journal of Computer Assisted Radiology and Surgery, 14*(10), 1673-1686. doi:10.1007/s11548-019-01927-z

Sethi, T., & Kantardzic, M. (2018). Data-driven exploratory attacks on black-box classifiers in adversarial. *Neurocomputing, 289*, 129-143. doi:10.1016/j.neucom.2018.02.007

Thomas, R. (2019). The new era of NLP. *Scientific Computing with Python.* Austin, Texas: SciPy. Retrieved from https://youtu.be/KChtdexd5Jo

Upchurch, M. (2018). Robots and AI at work: the prospects for singularity. *New Technology, 33*(3), 205-218. doi:10.1111/ntwe.12124

Wildberger, A. (1996). Introduction and overview of artificial life evolving intelligent agents for modeling and simulation. *Winter Simulation Conference* (pp. 161-168). doi:10.1109/WSC.1996.873274