Tae Wan Kim: Artificial Intelligence in the workplace through an ethical lens

Elva A. Resendez, PhD

Abstract
Artificial intelligence today continues to assimilate into our modern service industry occupations such as business, economics and health care. It is predicted that AI will continue to develop exponentially as a business staple and will follow a similar path to The Internet and Social Media in terms of growth opportunities and also the criticality to business. The concern for many, including Dr. Tae Wan Kim, associate professor of ethics at Carnegie Mellon University, is how the ethics of AI will affect individuals and society. In his recent book review, *Machines Like Me,* Kim writes on the varied ethical systems potentially used in AI training as he discusses the character development of the text. Whether a deontological approach or a consequentialist approach, AI is only as good as the training and the data provided. In this interview he says that to really address ethics issues in AI, much more funding should be given to moral philosophy and applied ethics.

ETHICS

EAR: What do you consider to be the greatest ethical violation of AI (present or future)?

TWK: A serious concern of mine is to "*trust AI without evidence*". Most humans do not understand how AI is developed/trained and therefore are naïve about AI. People tend to believe AI is smarter than humans, infallible or believe AI is more capable of doing the right thing or being fair. AI is trained by people; therefore, its actions, outcomes and fairness are similar to what has been done in the past by people.

EAR: Ethical issues exist in every field and machine learning is only beginning to adapt. How would you suggest those involved with AI and machine learning prepare their own ethical compass? What activities would you suggest to ensure individuals are in the "right" headspace?

TWK: People are more interested in ethics today than ever. Perhaps, it's a golden age for ethics. Technological innovation helps people to reconsider what is "right". The uncommon combination of tech and ethics helps people to consider ethics in a robust manner. To program an AI system with ethical rules, you should clarify the rules first and test whether the rules are good enough. To know whether the rules are good enough, you should use the various ethics tests moral philosophers have articulated for decades. Ethics is complex. There is no simple answer or short-cut. There is no easy framework or mindset to solve everything. Just as you take several courses to learn financial investment, you should spend as much time studying ethics. Many parts of ethical issues are philosophically technical (Kim, Hooker & Donaldson, Forthcoming), similarly, many parts of computer science issues are technical. As society relies on computer

scientists for technical issues, people can get help from moral philosophers. People who talk about AI ethics and newspaper articles are worried about it, yet funding to evaluate moral philosophy or applied ethics has not been accordingly increased. If we want to really address ethics issues in AI, much more funding should be given to moral philosophy and applied ethics.

BUSINESS

ER: Concerning AI in the workplace, what is your most pressing concern?

TWK: My most pressing concerns revolve around unemployment, the meaningfulness of work and how AI may influence an individual's purpose in the workplace. When an individual retires, like my father who retired, they may have money, family and activities, yet, they seek to fill a meaningful purpose previously fulfilled by work. Should AI replace workers, what will workers do to seek meaning in society? Do people need to work? Will they retrain for other work? If people retrain for other work, would it be faster to train an AI versus a person? Will rates of depression rise due to AI replacement? What would happen to work relationships in a 20% AI, 80% human workforce? Basically, business scholars need to examine the relationships tied to meaningfulness of work and the role of AI (Kim & Scheller-Wolf, 2019).

ER: In your book review of Machines Like Me (Kim, Forthcoming), you mentioned the not-so-rare potential of AI for gender discrimination in black box models as machine learning develops. What is the potential role of a "fairness" algorithm help to resolve the issues of workplace discrimination and with what other classifiers, race, age, etc.?

TWK: A growing number of people are studying fairness algorithm. But the field is not yet mature to answer your question. We do not yet know whether any fairness algorithms really can solve algorithmic biases. We do not yet know whether the statistical fairness approach, the mainstream approach (see Corbett-Davies & Goel, 2018 for a survey), is the right method. The most fundamental cause of algorithmic bias is that we are biased and AI's training data comes from us. If we can have un-biased data, most of the problem will be solved, unless developers intentionally attempt to discriminate against chosen groups. But it is difficult to identify whether training data has biases by looking at the data itself. For instance, Amazon's recruitment system did not consider protected class attributes (race, gender, disability etc.) and any proxies for them. This technique is called "anti-classification." But the system did not work as intended. Amazon quickly realized that the system preferred men to women significantly and the company had to shut down the system. The training data already had hid biases against women, and the system gave lower scores to any patterns that the machines categorized as feminine. The system by itself found hidden proxies about the patterns. In short, even if protected attributes are removed upfront, biases embedded in the training data as a network will be learned by the machine. There are other approaches such as statistical parity, but none of the solutions on the table are proved to be good enough. My proposal is that we should do what we can do first. First, develop machines to be well aligned with discrimination laws, before trying to align it with philosophers'

competing ideas about fairness. Second, we should not think that we can have a perfect notion of fairness that can be learnt by a machine without any controversy. Instead, we should be open to a piecemeal process. Various stakeholders' inputs, which are often inconsistent with each other, should be balanced through trial-and-error processes. Simultaneously, there must be a retrospective grievance system in which damages, if any, are repaired, victims are saved; wrongs are righted, and wrongdoers are punished, case by case.

ER: Much of social media relies upon tracking user patterns. In Sentiment Analysis, what is the real danger of manipulation by external sources (hackers, social media platforms, paid-for services) to skew real data for use by business?

TWK: Sentiment Analysis can be used to manipulate stock market in theory, but sentiment analysis has not yet been proven to be useful enough. Anyway, let me explain how such manipulation is possible. First, stock markets do not reflect actual markets but only what people think markets are like. Second, Sentiment Analysis predicts patterns between stock prices and sentiments. Third, companies can influence media sentiments by manipulating newspaper articles or social network services. Then, companies can manipulate stock prices by manipulating media sentiments. It's a possible scenario.

SOCIETY

EAR: In consideration of our current pandemic state, bots can safely perform some tasks and this is leading to loss of jobs in some industries. Does technological innovation justify mass unemployment?

TWK: This is a big question. First, we need more studies, both theoretical and empirical, to know whether innovation in AI technologies will really lead to mass unemployment and if so, when and how much. There are well written papers out there (e.g., Acemoglu, Autor, Hazell & Restrepo, 2020). But we still don't have a complete picture. The question has several layers. First, we need to predict when AI will outperform humans and in which domain. Second, the fact that AI outperforms humans in a certain domain does not by itself means that humans will be replaced by AI in that domain. If it's less costly to hire humans than AI, companies have no good reason to automate. Automation is often costly. Even if it's less costly to hire AI, there are other things to consider, such as, impact upon society. Just imagine an extreme case in which companies significantly automate their operations so that mass unemployment really occurs. In that case, customers have decreased purchasing power because they are not employed. Then, automation backfires companies' original goal to maximize profits. An alternative is a world in which a government gives something like basic income to everyone so that the unemployed can maintain purchasing power. In this world, perhaps, some of the unemployed may be very happy because they don't have to work; while some of them may be unhappy, because they have difficulty finding meaning without employment. It is not easy to predict the future. But we as

society should be precautious about various possibilities. Back to the question, technical innovation per se does not justify mass unemployment.

EAR: Do you agree with Stephen Hawking's statement that robots spell doomsday for mankind?

TWK: I don't know. To explain why let me use a common distinction: weak vs. strong AI. Strong AI is also called GAI (General Artificial Intelligence; or AGI). Strong AI/General AI is what you often see in futuristic movies. An example of general intelligence is a human who can solve problems across domains. Weak AI is a domain-specific system. Most current AI is considered weak AI. For example, an AI deployment trained for language translation is not able to drive a vehicle. Human translators can drive a car. There are researchers specifically dedicated to studying and developing GAI but realizing GAI *is not within our reach*. So, it's even difficult for me to imagine GAI. According to a recent survey, 352 top AI researchers predicted that domain-specific AI will outperform humans in the next decade in various domains including translating languages (by 2024), writing high-school essays (by 2026), driving a truck (by 2027), and working in retail (by 2031); but the same researchers predicted that there is only a 50% chance of automating all human jobs in 120 years (Grace, Salvatier, Dafoe, Zhang & Evans, 2018). Furthermore, the prediction concerns only domain-specific AI. Developing a single AI system that can automate all human jobs simultaneously is a totally different thing. For the sake of reality, my work is limited to weak AI. The so-called "superintelligence" (Alfonseca et al., 2021) is far beyond the realization of GAI, so I do not discuss it either. To answer your question, I am agnostic to Hawking's thesis.

ER: What do you feel is the next big trend in AI development for consumers to be aware of with the potential to affect their daily lives?

TWK: There is not much room for innovation in standard neural nets-based machine learning, except for its application to specific problems. But with more training data created by user activities, algorithms will improve. So, for instance, targeted advertising will get more sophisticated and individualized. In contrast, there are researchers who want to make a fundamental breakthrough in AI. The current model is a bunch of correlations and lacks causality. There is an attempt to develop a causal AI model. Some people say it's an oxymoron. But having a causal AI model will be useful to address some ethical issues. For instance, with a causal model, we can better identify liability/responsibility issues and with the counter-factual function, causal AI can be inherently explainable. Another good trend is the attempt to develop "neuro-symbolic AI." Like I said, all the problems we face are from neural nets-based AI and especially its lack of high-level cognitive functions such as rule-based reasoning, logical thinking, etc. There is a kind of AI that can do the high-level functioning, which is called "symbolic AI." Companies like IBM are working hard to develop to combine neural nets with symbolic AI, with which, we can address many of the ethics issues we face with standard machine learning.

ER: What advice would you give to practitioners or scholars interested in keeping up with AI development? What or where should they be watching?

TWK: People should learn more on how to train AI /a machine learning system. By watching/reading, people can quickly see AI limitations. People can see how fragile machine learning is, especially when it is used slightly outside the training set. Attaching a small black tape to a traffic light can make an autonomous vehicle go crazy. Just changing a few pixels in an image can cause similar misclassification. Where can people watch all of these developments? It's everywhere. In newspapers, YouTube, and online courses.

REFERENCES

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2020). AI and Jobs: Evidence from Online Vacancies (No. w28257). National Bureau of Economic Research.

Alfonseca, M., Cebrian, M., Fernández-Anta, A., Coviello, L., Abeliuk, A. & Rahwan, I. (2016). Superintelligence Cannot be Contained: Lessons from Computability Theory. Journal of Artificial Intelligence Research. 70. 10.1613/jair.1.12202.

Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. ArXiv, abs/1808.00023.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When Will AI Exceed Human Performance? Evidence from AI Experts. Journal of Artificial Intelligence Research 62 (2018): 729-754.

Kim, T.W. "Flawed like us and the starry moral law: a review of Ian McEwan's Machines Like Me", Journal of Business Ethics (Forthcoming).

Kim, T.W., Hooker, J., & Donaldson, T. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. Journal of Artificial Intelligence Research (Forthcoming).

Kim, T.W., & Scheller-Wolf, A. (2019). Technological unemployment, meaning in life, purpose of business, and the future of stakeholders. Journal of Business Ethics 160.2: 319-337.