Artificial Intelligence and Ethics

Ten areas of interest

Brian Patrick Green

Brian Green is the assistant director of Campus Ethics at the Markkula Center for Applied Ethics. Views are his own.

Artificial intelligence and machine learning technologies are rapidly transforming society and will almost certainly continue to do so in the coming decades. This social transformation will have deep ethical impact, with these powerful new technologies both improving and disrupting human lives. AI, as the externalization of human intelligence, offers us in amplified form everything that humanity already is, both good and evil. Much is at stake. At this crossroads in history we should think very carefully about how to make this transition, or we risk empowering the grimmer side of our nature, rather than the brighter.

The Markkula Center for Applied Ethics recently joined the Partnership on AI to Benefit People and Society, and as an institution we have been thinking deeply about the ethics of AI for several years. In that spirit, we offer a preliminary list of issues with ethical relevance in AI and machine learning.

1. Technical Safety (failure, hacking, etc.)

The first question for any technology is whether it works as intended. Will AI systems work as they are promised or will they fail? If and when they fail, what will be the results of those failures? And if we are dependent upon them, will we be able to survive without them?

For example, at least one person has already died in a semi-autonomous car accident because the vehicle encountered a situation in which the manufacturer anticipated it would fail and expected the human driver to take over, but the human driver didn't correct the situation.

The question of technical safety and failure is separate from the question of how a properly-functioning technology might be used for good or for evil (questions 3 and 4, below). This question is merely one of function, yet it is the foundation upon which all the rest of the analysis must build.

2. Transparency and Privacy

Once we have determined that the technology functions adequately, can we actually understand how it works and properly gather data on its functioning? Ethical analysis always depends on getting the facts first – only then can evaluation begin.

It turns out that with some machine learning techniques such as deep learning in neural networks it can be difficult or impossible to really understand why the machine is making the choices that it makes. In other cases, it might be that the machine can explain something, but the explanation is too complex for humans to understand.

Ethics Resources

Using This Site
Ethics App
Ethical Decision Making
Ethics Articles
Ethics Blogs
Ethics Cases
Ethics Curricula
Ethics Links
Ethics Podcasts
Ethics Spotlight
Ethics Training
Ethics Videos

OTHER CENTER BLOGS

- Benison: The Practice of Ethical Leadership
- Internet Ethics: Views from Silicon Valley
- Ethical Dilemmas in the Social Sector

For example, in 2014 a computer proved a mathematical theorem called the "Erdos discrepancy problem," using a proof that was, at the time at least, longer than the entire Wikipedia encyclopedia. Explanations of this sort might be true explanations, but humans will never know for sure.

As an additional point, in general, the more powerful someone or something is, the more transparent it ought to be, while the weaker someone is, the more right to privacy he or she should have. Therefore the idea that powerful AIs might be intrinsically opaque is disconcerting.

3. Malicious Use & Capacity for Evil

A perfectly well functioning technology, such as a nuclear weapon, can, when put to its intended use, cause immense evil. Artificial intelligence, like human intelligence, will be used maliciously, there is no doubt.

For example, AI-powered surveillance is already widespread, in both appropriate contexts (e.g., airport-security cameras) and perhaps inappropriate ones (e.g., products with always-on microphones in our homes). More obviously nefarious examples might include AI-assisted computer-hacking or lethal autonomous weapons systems (LAWS), a.k.a. "killer robots." Additional fears, of varying degrees of plausibility, include scenarios like those in the movies "2001: A Space Odyssey," "Wargames," and "Terminator."

While movies and weapons technologies might seem to be extreme examples of how AI might empower evil, we should remember that competition and war are always primary drivers of technological advance, and that militaries and corporations are working on these technologies right now. History also shows that great evils are not always completely intended (e.g., stumbling into World War I and various nuclear close-calls in the Cold War), and so having destructive power, even if not intending to use it, still risks catastrophe. Because of this, forbidding, banning, and relinquishing certain types of technology would be the most prudent solution.

4. Beneficial Use & Capacity for Good

The main purpose of AI is, like every other technology, to help people lead longer, more flourishing, more fulfilling lives. This is good, and therefore insofar as AI helps people in these ways, we can be glad and appreciate the benefits it gives to us.

Additional intelligence will likely provide improvements in nearly every field of human endeavor, including, for example, archaeology, biomedical research, communication, data analytics, education, energy efficiency, environmental protection, farming, finance, legal services, medical diagnostics, resource management, space exploration, transportation, waste management, and so on.

As just one concrete example of a benefit from AI, some farm equipment now has computer systems capable of visually identifying weeds and spraying them with tiny targeted doses of herbicide. This not only protects the environment by reducing the use of chemicals on crops, but it also protects human health by reducing exposure to these chemicals.

5. Bias in Data, Training Sets, etc.

One of the interesting things about neural networks, the current workhorses of artificial intelligence, is that they effectively merge a computer program with the data that is given to it. This has many benefits, but it also risks biasing the entire system in unexpected and potentially detrimental ways.

Already algorithmic bias has been discovered, for example, in areas ranging from criminal sentencing to photograph captioning. These biases are more than just embarrassing to the corporations which produce these defective products; they have concrete negative and harmful effects on the people who are victims of these biases, as well as reducing trust in corporations, government, and other institutions which might be using these biased products. Algorithmic bias is one of the major concerns in AI right now and will remain so in the future unless we endeavor to make our technological products better than we are. As one person said at a recent meeting of the Partnership on AI, "We will reproduce all of our human faults in artificial form unless we strive right now to make sure that we don't."

6. Unemployment / Lack of Purpose & Meaning

Many people have already perceived that AI will be a threat to certain categories of jobs. Indeed, automation of industry has been a major contributing factor in job losses since the beginning of the industrial revolution. AI will simply extend this trend to more fields, including fields that have been traditionally thought of as being safer from automation, for example law, medicine, and education. Other than the job of AI developer, it is not clear what new careers these unemployed people will be able to transition into, and even AI programming may become at least partially automated in the future.

Attached to the concern for employment is the concern for how humanity spends its time and what makes a life well-spent. What will millions of unemployed people do? What good purposes can they have? What can they contribute to the well-being of society? How will society prevent them from becoming disillusioned, bitter, and swept up in evil movements such as white supremacy and terrorism?

7. Growing Socio-Economic Inequality

Related to the unemployment problem is the question of how people will survive if unemployment rises to very high levels. Where will they get money to maintain themselves and their families? While prices may decrease due to lowered cost of production, those who control AI will also likely rake in much of the money that would have otherwise gone into the wages of the now-unemployed, and therefore economic inequality will increase.

Some people, including some billionaires like Mark Zuckerberg, have suggested a universal basic income (UBI) to address the problem, but this will require a major reconstruction of national economies. Various other solutions to this problem may be possible, but they all involve potentially major changes to human society and government. Ultimately this is a political problem, not a technical one, so this solution, like those to many of the problems described here, needs to be addressed at the political level.

8. Moral De-Skilling & Debility

If we turn over our decision-making capacities to machines, we will become less experienced at making decisions. For example, this is a well-known phenomenon among airline pilots: the autopilot can do everything about flying an airplane, from take-off to landing, but pilots intentionally choose to manually control the aircraft at crucial times in order to maintain their piloting skills.

Because one of the uses of AI will be to either assist or replace humans at making certain types of decisions (e.g. spelling, driving, stock-trading, etc.), we should be aware that humans may become worse at these skills. In its most extreme form, if AI starts to make ethical and political decisions for us, we will become worse at ethics and politics. We may reduce or stunt our moral development precisely at the time when our power has become greatest and our decisions the most important.

This means that the study of ethics and ethics training are now more important than ever. We should determine ways in which AI can actually enhance our ethical learning and training. We should never allow ourselves to become de-skilled and debilitated at ethics, or when our technology finally does present us with a problem we must solve we may be like frightened and confused children before a creation we do not understand.

9. AI Personhood / “Robot Rights”

Some thinkers have wondered whether AIs might eventually become self-conscious, attain their own volition, or otherwise deserve recognition as persons like ourselves. Legally speaking, personhood has been given to corporations and (in other countries) rivers, so there is certainly no need for consciousness even before legal questions may arise.

Morally speaking, we can anticipate that technologists will attempt to make the most human-like AIs and robots possible, and perhaps someday they will be such good imitations that we will wonder if they might be conscious and deserve rights – and we might not be able to determine this conclusively. If future humans do conclude AIs and robots might be worthy of moral status, then we ought to err on the side of caution and fairness and give it.

In the midst of this uncertainty about the status of our creations, what we will know, though, is that *we humans* have moral characters and that, to quote Aristotle, "we become what we repeatedly do." So we ought not to treat AIs and robots badly, or we might be habituating ourselves towards having flawed characters, regardless of the moral status of the artificial beings we are interacting with. In other words, no matter the status of AIs and robots, for the sake of our own moral characters we ought to treat them well, or at least not abuse them.

10. Effects on the Human Spirit

All of the above areas of interest will have effects on how humans perceive themselves, relate to each other, and live their lives. But there is a more existential question too. If the purpose and identity of humanity has something to do with our intelligence (as several prominent Greek philosophers believed, for example), then by externalizing our intelligence and improving beyond human intelligence, are we making ourselves second-class beings to our own creations?

This is a deeper question with artificial intelligence which cuts to the core of our humanity, into areas traditionally reserved for philosophy, spirituality, and religion. What will happen to the human spirit if or when we are bested by our own creations in everything that we do? Will human life lose meaning? Will we come to a new discovery of our identity beyond our intelligence? Perhaps intelligence is not as important to our identity as we might think it is, and perhaps turning over intelligence to machines will help us to realize that.

This is just a start at the exploration of the ethics of AI; there is much more to say. New technologies are always created for the sake of something good – and AI offers us amazing new powers. Through the concerted effort of many individuals and organizations, we can hope to use AI to make a better world.

This article is an adaptation of a paper presented to the Pacific Coast Theological Society, November 3rd, 2017. A shorter draft was presented on October 24th, 2017, at Santa Clara University at a panel entitled "AI: Ethical Challenges and a Fast Approaching Future." In the panel, I presented a list of nine areas of ethical concern; thanks to some helpful feedback I expanded the list to ten.

Nov 21, 2017