INTRO TO ETHICS – TOPIC 3

R. D. Walsh, Ph.D.

Table of Contents

TOPIC 3
ARTIFICIAL INTELLIGENCE ETHICS

Introduction

The purpose of this Topic is to investigate, from an ethical perspective, the burgeoning development and deployment in everyday human life of the various technologies that comprise the field of Artificial Intelligence (AI), especially AI technology in its pursuit of a super-intelligent, completely autonomous, self-learning, and perhaps self-replicating machine. What will be the impact of such artificial, mechanically configured, morality-free superintelligence on human life? What moral concerns accompany the development of super AI?

It is clear to anyone who takes even a cursory look that the field of AI that it is growing with exponential speed in many different directions. What is happening is a "cognitization" of everyday life, from smart phones to smart cars, houses, and cities. Life-changing technological breakthroughs in AI, robotics, and associated fields, seem to happen on an almost daily basis. The future seems wide open. In fact, the future is so wide open that it is very difficult to see what

might be coming down the road in future years, or even what is just up ahead a few months. That feels a little scary to some people.

Numerous researchers, such as Nick Bostrom, author of the book ***Superintelligence***, think that the development of super-intelligent (having *general* intelligence), self-learning, and, ultimately, autonomous machines—far surpassing human abilities (which machines already do now in some 'narrow' areas)—is something that could happen with unexpected speed. And while such development may have many attractive benefits for humans, Bostrom is among those who believe that this also should be cause for concern and advance planning now—while there's still time. Wide-open futures have high existential risk. Now is the time to consider moral parameters and advance directives before AI gets out of control. Could it really get out of control?

Future uncertainty about the rapid development of AI with unforeseeable, even unimaginable outcomes is causing existential anxiety now. Indeed, voices have been raised in alarm. Elon Musk, while supporting research into the development of friendly AI, likens AI development to "summoning the demon" which he believes *will* inevitably get out of control and perhaps turn on its creator unless precautions are taken in advance. Stephen Hawking, the well-known physicist, thinks AI could cause the end of the human race. Bill Gates doesn't understand "why some people are not concerned" at all about AI. Wake up, folks! Gates wants to say. While it is true that AI and robotics has already made many positive contributions that have benefitted human beings, there are numerous sources of moral concern.

Urvashi Aneja points out a number of the potential areas of moral concern with the development of AI in her article below entitled, "Why our conversations on Artificial Intelligence are Incomplete." Aneja believes that the public should be more aware of the potential negative impact of AI and that in addition to all the excitement about what AI can practically accomplish, there should be a broader discussion about what it should, *and should not,* accomplish, and what values should guide that development. To what end should AI be directed? Profit? Social benefit? Both? In what proportion? How should that be accomplished? Is self-regulation by the tech industry sufficient? How should the benefits from the AI revolution be distributed? To what extent should robots be held accountable for their behavior? What impact will AI have on moral agency? On society as a whole?

One important area of AI concern that Aneja does not focus on specifically is the development of AI autonomous weapons systems and the catastrophic possible outcomes from an all-out AI arms race. There have been numerous calls for limits to the development of autonomous weapons, such as the following from the Future of Life Institute:

> Unlike nuclear weapons, [autonomous weapons] require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and

> selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity.[1]

Human Intelligence, Artificial Intelligence, and Multiple Intelligences

"Artificial" means something created to look like or simulate something real although it is not the real thing itself. Artificial flowers look more or less *like* real flowers but they are not real (natural, organic) flowers. The same for artificial intelligence in relation to human intelligence.

Artificial intelligence attempts to mimic or simulate human intelligence. It produces output to appear like the kind of output achieved by human intelligence. In this connection, Sam Harris has asserted from an empirical perspective that "intelligence" is basically a matter of information processing, and that information or data processing is essentially what the human brain does; and this can be simulated by machines. But is that really the case? It seems like it is the case, at least in narrow applications such as scanning large amounts of data for clusters of similarities. Can human intelligence be effectively summed up as a process of neural data processing? I wonder about that.

Isn't human cognition in all of its many and varied manifestations in science, art, business, entertainment, literature, technology, architecture, industry, and in every area of human endeavor … isn't this creative intelligence ultimately beyond simulation? Will it ever be possible to effectively reduce human cognition (in all of its non-rational and inexplicable spontaneity and multiplicity) to so-called self-learning data-processing machines? That seems like a narrow stance on the idea of human intelligence. In many ways machine intelligence may be and become more effective than human intelligence, and human intelligence will give way to this convenience, simplicity, and efficiency, because that's what humans do.

Certainly, simulated intelligence is itself a *kind* of intelligence, but hardly the only kind. That there are many kinds of intelligence was articulated by Howard Gardner who introduced the concept of "multiple intelligences" fifty years ago. He argued that we should change our educational models to fit these different types of intelligence. Can all of these multiple intelligences be simulated by machines?

Reducing general, cross-domain human, embodied intelligence to a singular function of data or information processing (which machines already do far better than humans in narrow applications) is a huge oversimplification. To represent this oversimplified intelligence as effectively simulating human intelligence is to do conceptual damage to natural human intelligence.

Human intelligence is an infinitely complex, non-linear, intuitively-driven, mostly tacit, insightful rather than strictly logical process that is ultimately inexplicable even to itself, mysterious, and never-endingly adaptive and creative. This cannot be captured in any two-dimensional definitional straightjacket. The damage to human intelligence ultimately may be in

---

[1] Future of Life Institute. www.futureoflife.org July 28, 2015. https://futureoflife.org/open-letter-autonomous-weapons/

the form of humans sacrificing their infinitely complex but somewhat messy multiplicity of overlapping and interweaving natural human intelligence to the cold efficiency of super-intelligent calculated output. Thus, we might increase our leisure by sacrificing our birthright.

What potential dangers for human beings might there be coming along the unforseeable path of this intelligence explosion?

Some AI questions raised by the following article

1. How will AI impact the labor market? Jobs? The capital market? Overall, will AI have a positive or negative value on future society? Short term? Long term?

2. Should AI be developed/deployed strictly for profit? What about social/political benefits?

3. How should the benefits of AI be distributed? On what principle should the mechanism of distribution of AI benefits and burdens be based?

4. Will self-regulation of AI deployment by tech industry work?

5. Should there be greater algorithmic transparency? What about data bias and algorithmic bias of protected characteristics like race and gender?

6. Algorithmic transparency versus competitive advantage? AI power must not be invisible

7. Will AI intensify existing social injustices, like racial discrimination and discrimination against women?

8. Will robots have agency? Moral/legal responsibility? Autonomy possible?

9. The runaway trolley reappears…What would a robot do?

10. AI profitability must be tied to questions of purpose, values, accountability

11. Need expansion of AI conversation beyond AI "epistemic community"

Aneja: Why our conversations on AI are incomplete[2]

by Urvashi Aneja

Dr. Urvashi Aneja

Artificial Intelligence (AI) is no longer the subject of science fiction and is profoundly transforming our daily lives. While computers have already been mimicking human intelligence for some decades now using logic and if-then kinds of rules, massive increases in computational power are now facilitating the creation of 'deep learning' machines i.e. algorithms that permit software to train itself to recognize patterns and perform tasks, like speech and image recognition, through exposure to vast amounts of data.

These deep learning algorithms are everywhere, shaping our preferences and behavior Facebook uses a set of algorithms to tailor what news stories an individual user sees and in what order. Bot activity on Twitter suppressed a protest against Mexico's now president by overloading the hashtag used to organize the event. The world's largest hedge fund is building a piece of software to automate the day-to-day management of the firm, including, hiring, firing and other strategic decision-making. Wealth management firms are increasingly using algorithms to decide where to invest money. The practice of traders shouting and using hand signals to buy and sell commodities has become outdated on Wall Street as traders have been replaced by machines. And bots are now being used to analyze legal documents to point out potential risks and areas of improvement.

Much of the discussion on AI in popular media has been through the prism of job displacement. Analysts, however, differ widely on the projected impact – a 2016 study by the Organization for Economic Co-operation and Development estimates that 9% of jobs will be diplaced in the next two years, whereas a 2013 study by Oxford University estimates that job displacement will be 47%. The staggering difference illustrates how much the impact of AI remains speculative.

---

[2] Aneja, Urvashi. "Why our conversations on AI are incomplete." *The Wire*. February 19, 2017. https://thewire.in/109882/why-our-conversations-on-artificial-intelligence-are-incomplete/

Responding to the threat of automation on jobs will undoubtedly require revising existing education and skilling paradigms, but at present, we also need to consider more fundamental questions about the purposes, values and accountability of AI machines. Interrogating these first-order concerns will eventually allow for a more systematic and systemic response to the job displacement challenge as well.

First, what purpose do we want to direct AI technologies towards? AI technologies can undoubtedly create tremendous productivity and efficiency gains. AI might also allow us to solve some of the most complex problems of our time. But we need to make political and social choices about the parts of human life in which we want to introduce these technologies, at what cost and to what end.

Technological advancement has resulted in a growth in national incomes and GDP, yet the share of national incomes that have gone to labour has dropped in developing countries. Productivity and efficiency gains are thus not in themselves conclusive indicators on where to deploy AI – rather, we need to consider the distribution of these gains. Productivity gains are also not equally beneficial to all – incumbents with data and computational power will be able to use AI to gain insight and market advantage.

Moreover, a bot might be able to make more accurate judgments about worker performance and future employability, but we need to have a more precise handle over the problem that is being addressed by such improved accuracy. AI might be able to harness the power of big data to address complex social problems. Arguably, however, our inability to address these problems has not been a result of incomplete data – for a number of decades now we have had enough data to make reasonable estimates about the appropriate course of action. It is the lack of political will and social and cultural behavioural patterns that have posed obstacles to action, not the lack of data. The purpose of AI in human life must not be merely assumed as obvious, or subsumed under the banner of innovation, but be seen as involving complex social choices that must be steered through political deliberations.

This then leads to a second question about the governance of AI – who should decide where AI is deployed, how should these decisions be made and on what principles and priorities? Technology companies, particularly those that have the capital to make investments in AI capacities, are leading current discussions predominantly. Eric Horvitz, managing director of the Microsoft Research Lab, launched the One Hundred Year Study on Artificial Intelligence based out of Stanford University. The Stanford report makes the case for industry self-regulation, arguing

that 'attempts to regulate AI, in general, would be misguided as there is no clear definition of AI and the risks and considerations are very different in different domains.'

The White House Office of Science and Technology Policy recently released a report on the 'Preparing for the Future of Artificial Intelligence', but accorded a minimal role to the government as regulator. Rather, the question of governance is left to the supposed ideal of innovation – i.e. AI will fuel innovation, which will fuel economic growth and this will eventually benefit society as well. The trouble with such innovation-fueled self-regulation is that development of AI will be concentrated in those areas in which there is a market opportunity, not necessarily areas that are the most socially beneficial. Technology companies are not required to consider issues of long-term planning and the sharing of social benefits, nor can they be held politically and socially accountable.

Earlier this year, a set of principles for Beneficial AI was articulated at the Asilomar Conference – the star speakers and panelists were predominantly from large technology companies like Google, Facebook and Tesla, alongside a few notable scientists, economists and philosophers. Notably missing from the list of speakers was the government, journalists and the public and their concerns. The principles make all the right points, clustering around the ideas of "beneficial intelligence", "alignment with human values" and "common good", but they rest on fundamentally tenuous value questions about what constitutes human benefit – a question that demands much wider and inclusive deliberation, and one that must be led by government for reasons of democratic accountability and representativeness.

What is noteworthy about the White House Report in this regard is the attempt to craft a public deliberative process – the report followed five public workshops and an Official Request for Information on AI.

The trouble is not only that most of these conversations about the ethics of AI are being led by the technology companies themselves, but also that governments and citizens in the developing world are yet to start such deliberations – they are in some sense the passive recipients of technologies that are being developed in specific geographies but deployed globally. The Stanford report, for example, attempts to define the issues that citizens of a typical North American city will face in computers and robotic systems that mimic human capabilities. Surely these concerns will look very different across much of the globe. The conversation in India has mostly been clustered around issues of jobs and the need for spurring AI-based innovation to

accelerate growth and safeguard strategic interests, with almost no public deliberation around broader societal choices.

The concentration of an AI epistemic community in certain geographies and demographics leads to a third key question about how artificially intelligent machines learn and make decisions. As AI becomes involved in high-stakes decision-making, we need to understand the processes by which such decision making takes place. AI consists of a set of complex algorithms built on data sets. These algorithms will tend to reflect the characteristics of the data that they are fed. This then means that inaccurate or incomplete data sets can also result in biased decision making. Such data bias can occur in two ways.

First, if the data set is flawed or inaccurately reflects the reality it is supposed to represent. If for example, a system is trained on photos of people that are predominantly white, it will have a harder time recognizing non-white people. This kind of data bias is what led a Google application to tag black people as gorillas or the Nikon camera software to misread Asian people as blinking. Second, if the process being measured through data collection itself reflects long-standing structural inequality. *ProPublica* found, for example, that software that was being useful to assess the risk of recidivism in criminals was twice as likely to mistakenly flag black defendants as being at higher risk of committing future crimes. It was also twice as likely to incorrectly flag white defendants as low risk.

What these examples suggest is that AI systems can end up reproducing existing social bias and inequities, contributing towards the further systematic marginalization of certain sections of society. Moreover, these biases can be amplified as they are coded into seemly technical and neutral systems that penetrate across a diversity of daily social practices. It is, of course, an epistemic fallacy to assume that we can ever have complete data on any social or political phenomena or peoples. Yet, there is an urgent need to improve the quality and breadth of our data sets, as well as investigate any structural biases that might exist in these data – how we would do this is hard enough to imagine, leave alone implement.

The danger that AI will reflect and even exacerbate existing social inequities leads finally to the question of the agency and accountability of AI systems. Algorithms represent much more than code, as they exercise authority on behalf of organizations across various domains and have real and serious consequences in the analog world. However, the difficult question is whether this authority can be considered a form of agency that can be held accountable and culpable.

Recent studies suggest for example that algorithmic trading between banks was at least partly responsible for the financial crisis of 2008; the crash of the sterling in 2016 has similarly been linked to a panicky bot-spiral. Recently, both Google and Tesla's self-driving cars caused fatal crashes – in the Tesla case, a man died while using Tesla's autopilot function. Legal systems across the world are not yet equipped to respond to the issue of culpability in such cases, and the many more that we are yet to imagine. Neither is it clear how AI systems will respond to ethical conundrums like the famous trolley problem, nor the manner in which human-AI interaction on ethical questions will be influenced by cultural differences across societies or time. The question comes down to the legal liability of AI, whether it should be considered a subject or an object.

The trouble with speaking about accountability also stems from the fact that AI is intended to be a learning machine. It is this capacity to learn that marks the newness of the current technological era, and this capacity of learning that makes it possible to even speak of AI agency. Yet, machine learning is not a hard science; rather its outcomes are unpredictable and can only be fully known after the fact. This leads to an incompleteness problem for political and legal systems that are charged with the governance of AI.

The question of accountability also comes down to one of visibility. Any inherent bias in the data on which an AI machine is programmed is invisible and incomprehensible to most end users. This inability to review the data reduces the agency and capacity of individuals to resist, even recognize, the discriminatory practices that might result from AI. AI technologies thus exercise a form of invisible but pervasive power, which then also obscures the possible points or avenues for resistance. The challenge is to make this power visible and accessible. Companies responsible for these algorithms keep their formulas secret as proprietary information. However, the far-ranging impact of AI technologies necessitates the need for algorithmic transparency, even if it reduces the competitive advantage of companies developing these systems. A profit motive cannot be blindly prioritized if it comes at the expense of social justice and accountability.

When we talk about AI, we need to talk about jobs – both about the jobs that will be lost and the opportunities that will arise from innovation. But we must also tether these conversations to questions about the purpose, values, accountability and governance of AI. We need to think about the distribution of productivity and efficiency gains and broader questions of social benefit and well being. Given the various ways in which AI systems exercise power in social contexts, that power needs

to be made visible to facilitate conversations about accountability. And responses have to be calibrated through public engagement and democratic deliberation – the ethics and governance questions around AI cannot be left to market forces alone, albeit in the name of innovation.

Finally, there is a need to move beyond the universalizing discourse around technology – technologies will be deployed globally and with global impact, but the nature of that impact will be mediated through local political, legal, cultural and economic systems. There is an urgent need to expand the AI epistemic community beyond the specific geographies in which it is currently clustered, and provide resources and opportunities for broader and more diverse public engagement.

ABOUT DR. URVASHI ANEJA … Urvashi Aneja is Founding Director of Tandem Research, a multidisciplinary think tank based in Socorro, Goa that produces policy insights around issues of technology, sustainability and governance. She is Associate Professor at the Jindal School of International Affairs and Research Fellow at the Observer Research Foundation

Brolcháin: The battle for ethics at the cutting edge of technology[3]

Fiachra Brolcháin

Dr. Fiachra O. Brolchain

In an era of climate change, political instability, biodiversity loss and economic uncertainty, the pace of technological innovation is widely celebrated. Governments compete with each other to attract tech companies, with tax and education policies increasingly focused on the needs of technology developers. Some people speak of us being in the midst of a new Industrial Revolution. We seem to revere novel technologies and pin many of our hopes for the future upon them.

A large number of these technological developments bring many societal benefits, but our collective enthusiasm for technology can lead us to overlook or underplay many of the downsides. The speed of technological change – bringing us big data, driverless cars, genetic engineering and smart cities, with true AI and geo-engineering distinct future possibilities – is truly astounding. Society is like a jockey wearing a blindfold. The power and pace of the horse is exhilarating, but we have little to no idea where we are going.

That new technologies will significantly change our world is obvious. Whether this will be beneficial or harmful remains to be seen. Novel technologies and those in the early stages of development have the potential to exacerbate the myriad problems of the globe, or to mitigate them. Much will depend on the choices we make regarding their use.

These choices do not take place in a vacuum and ethical philosophy can provide us with guidance as we attempt to navigate our way. The choices available to us in relation to these new technologies are ethical choices. We need to be guided

[3] Brolcháin, Fiachra. "The battle for ethics at the cutting edge of technology." Siliconrepublic. March 21, 2017. www.siliconrepublic.com https://www.siliconrepublic.com/machines/ethics-technology-fiachra-o-brolchain-dcu

by our best ethical principles if we are to ensure that the current technological revolution does not result in misery for future generations.

Take, for instance, the burgeoning field of assistive technologies. A whole range of assistive technologies are now being developed to help people with physical or intellectual disabilities, as well as the ageing populations across the Western world. Addressing a range of needs, these tools are designed to make the lives of users and careers easier. These technologies will be used by the most vulnerable members of our society, making the ethical issues particularly important.

Indeed, the general populace is increasingly using assistive devices, from mobile phones to wearables. While there are clear benefits of assistive technologies, there are ethical concerns – the most prominent of which is a concern with privacy.

What do we mean when we talk about privacy? This is not an easy thing to answer. The meaning of privacy is historically and philosophically complex. Some argue that it is a moral right with inherent value; others contend that its value is instrumental.

Conceptually, privacy is often associated with human dignity and with the development of the authentic self. People are likely to behave differently when they know that they are being observed.

We need privacy if we are to avoid self-censorship, or if we are to be able to have certain discussions with each other. Without a space to think and explore various ideas, a person's psychological development is at risk of being stunted. This has led many thinkers to stress the normative importance of informational privacy – the idea that I should be able to control access to information about myself. Many of my thoughts, acts and words should be inaccessible to others. Novel technologies, including assistive technologies, that monitor and gather data about the person constitute a threat to privacy.

Why should we care about privacy? Privacy is also conceptually connected to the concept of autonomy, ie, being able to form your own opinions and make decisions without external influence. Autonomy is a central value in liberal thought, which reveres the liberty of the autonomous individual.

The autonomous individual weighs up their options, ponders their choices, and makes individual decisions without undue external influence. As new technologies – from big data to eye-tracking, facial recognition and emotion capture – undermine privacy, our autonomy is threatened. Increased data about the way

individuals are likely to behave, their preferences and dislikes, and their emotional responses to various stimuli, makes them easier to manipulate and control.

One might argue that those who don't want to share their information could simply refuse to use the new devices. However, this is unlikely to be sufficient. The internet of things – in which connected objects 'talk' to each other – promises the creation of 'smart cities'.

We will be living in cities where buildings can communicate with each other and with our devices, driverless cars will take us from place to place, and our fridges will remind us to buy milk. The benefits of these technologies have been heralded continuously and are, no doubt, real. For example, from an environmental perspective, increased data about air and water quality and energy use can play an important role in combatting climate change.

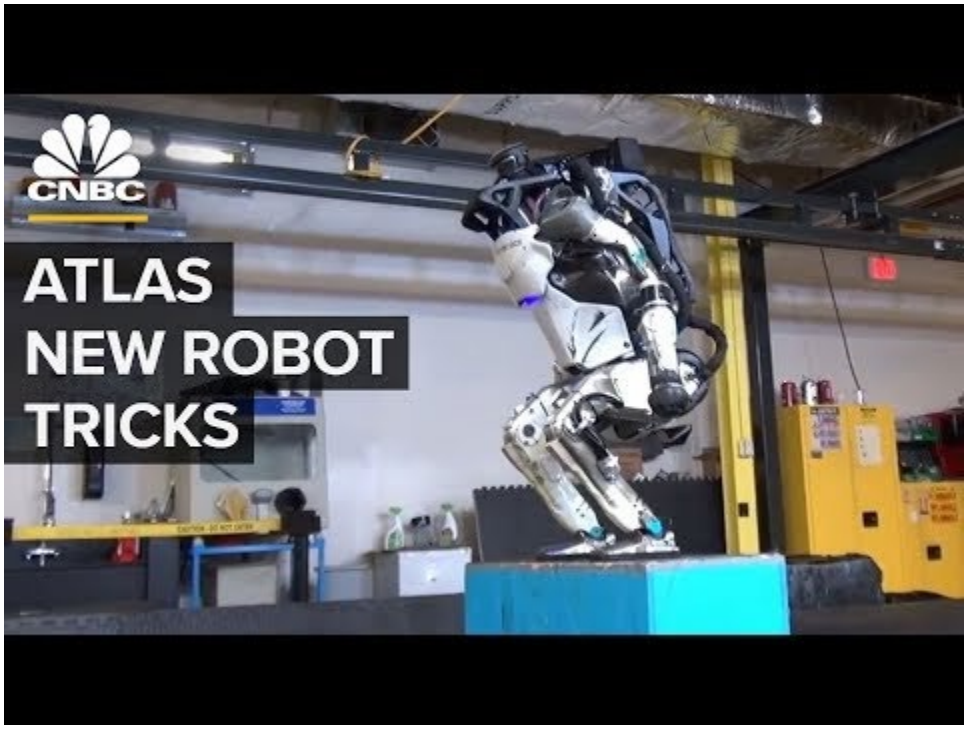
However, it will also mean that a person living in such a city could be continuously under surveillance. The use totalitarian regimes could make of such technologies would be familiar to Orwell.

Orwell's dystopian vision could yet be combined with that of Aldous Huxley's *Brave New World*. In a capitalist and consumerist society, much of the data about us will be used for commercial purposes. Omnipresent advertisers armed with huge data sets about each person would make it increasingly difficult for anyone to experience anything that has not been engineered and tailored to grab our individual attention. Already, our lives are inundated with demands on our attention – the internet of things and smart cities will exacerbate this while reducing our privacy significantly. Our mental lives will be less our own. Our encounters with the world will be mediated through technologies designed to catch our attention. This is far from the liberty and autonomy envisioned during the Enlightenment.

It is worth asking who will design these technologies and what their aims are. We must address the issue of responsibility for the negative impact of novel technologies. We must consider the reasons we hold for creating these new technologies – not just in terms of how they will benefit individual people and companies, but their overall societal effect.

The decisions we make now in relation to the technologies we are inventing will shape the societies we, and future generations, will live in. These choices will not take place in a moral vacuum and it is essential that we give deep consideration to the values guiding them.

ABOUT DR. BROLCHÁIN ….. Dr Fiachra Brolcháin has worked on various aspects of applied ethics, including the ethical and social implications of virtual reality and social networking in association with the EU's Reverie Project, and the ethical implications of human enhancement technologies. He is currently working as a Marie Curie ASSISTID Fellow at Dublin City University (DCU), looking at the ethics of the development, use and distribution of assistive technologies for people with intellectual disabilities and autism spectrum disorder.



Video: *Boston Robotics' Atlas has learned some new tricks (1:01)*

Nick Bostrom: The future of humanity

How do we invest in the future of humanity?

Swedish philosopher Nick Bostrom explains[4]

Economics correspondent Paul Solman recently traveled to Oxford University's Future of Humanity Institute. Solman spoke with the institute's founding director Nick Bostrom, a Swedish philosopher known for his work on artificial intelligence and existential threats. At the Future of Humanity Institute, Bostrom leads a team trying to figure out how to best invest in the future of humanity. That means identifying threats to the continuing existence of homo sapiens and figuring out how to reduce the possibility of such events.

Video: What happens when AI gets smarter than us? Dr. Nick Bostrom (16:30)

[4] Solman, Paul. "How do we invest in the future of humanity? Swedish philosopher Nick Bostrom explains." PBS NEWSHOUR/Making Sen$e. July 20, 2017. http://www.pbs.org/newshour/making-sense/invest-future-humanity-swedish-philosopher-nick-bostrom-explains/

PAUL SOLMAN: If I care about future generations, 100,000 years from now, and there's some possibility that they won't exist, what should I invest in to give them the best chance of survival and having a happy life the way I've had one?

NICK BOSTROM: What you should invest in is what we are trying to figure out, and it's a really difficult question. How can we trace out the links between actions that people take today and really long-term outcomes for humanity — outcomes that stretch out indefinitely into the future?

PAUL SOLMAN: And that's why [the institute] is called the Future of Humanity…

NICK BOSTROM: That's one of the reasons it's called that. So I call this effort macrostrategy — that is, to think about the really big strategic situation for having a positive impact on the long-term future. There's the butterfly effect: A small change in an initial condition could have arbitrarily large consequences. And it's hard enough to predict the economy two years from now, so how could we even begin to think about how your actions make a difference a million years from now? So there are some ideas that maybe bring the answer a little bit closer. One idea is this concept of existential risk. That helps focus our attention.

PAUL SOLMAN: Nuclear winter — that is, the period of abnormal cold that would follow a nuclear war. That has been, in my lifetime, I think the most common existential threat that people have talked about.

NICK BOSTROM: Well, if you think that nuclear war poses a threat to the survival of our species or even if you think that it would just be enormous destruction, then obviously we would look for ways to try to reduce the probability that there would be a nuclear war. So here you have to introduce a second consideration, which is how easy it is to actually make a difference to a particular race.

So it is quite difficult for some individual to reduce the probability of a nuclear war, because there are big nations with big stockpiles and strong incentives and a lot of money and a lot of people who have worked on this for decades. So if you, as an individual, choose to join a disarmament campaign, it might make some difference, but a small difference. So there might be other scenarios that have been more neglected and where maybe one extra person or one extra million dollars of research funding would make a larger, proportional difference. So you want to think, how big is the problem, and how much difference can you, on the margin, make to the degree to which the problem gets solved?

"So if there are big existential risks, I think they are going to come from our own activities and mostly from our own inventiveness and creativity."

PAUL SOLMAN: And one area that you yourself have been working on a lot is artificial intelligence, which you've called super intelligence. Is that an existential risk, do you think?

NICK BOSTROM: When I survey the possible things that could derail humanity's long-term future, it can roughly distinguish natural risks, such as volcano eruptions, earthquakes and asteroids, and risks that arise somewhere from our own activity. It's pretty clear that all the really big risks to our survival are of the latter kind, anthropogenic. We've survived risks from nature for 100,000 years, right? So, it's unlikely any of those things would do us in within the next 100 years. Whereas, in the next century, we will be inventing radical new technologies — machine intelligence, perhaps nanotech, great advances in synthetic biology and other things we haven't even thought of yet. And those new powers will unlock wonderful opportunities, but they might also bring with them certain risks. And we have no track record of surviving those risks. So if there are big existential risks, I think they are going to come from our own activities and mostly from our own inventiveness and creativity.

PAUL SOLMAN: What are the greatest of those risks?

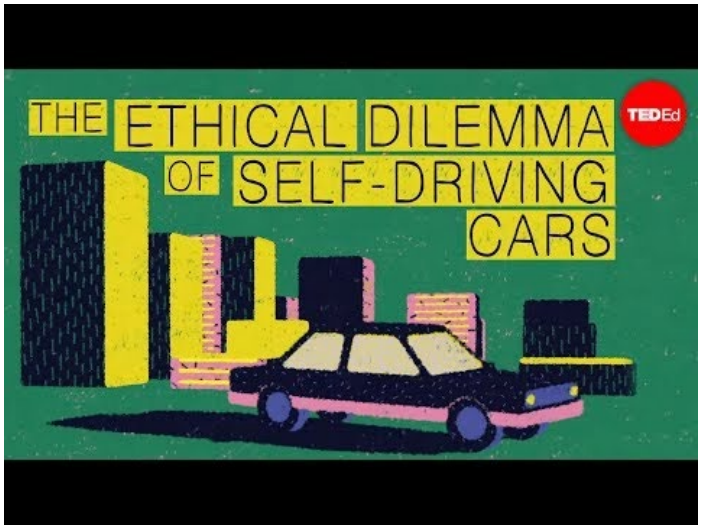
NICK BOSTROM: I think the greatest existential risks over the coming decades or century arise from certain, anticipated technological breakthroughs that we might make in particular, machine super intelligence, nanotechnology and synthetic biology. I think each of these has an enormous potential for improving the human condition by helping cure disease, poverty, etcetera. But one could imagine them being misused, used to create very powerful weapon systems, or even in some cases some kind of accidental destructive scenario, where we suddenly are in possession of some technology that's far more powerful than we are able to control or use wisely.

PAUL SOLMAN: How would you rank them in terms of the danger?

NICK BOSTROM: Biotech, synthetic biology and AI I think are near the top. I would also add the unknown. Suppose you had to ask me this question 100 years ago. What are the biggest existential risks? At that time, nobody would have mentioned AI; they didn't have computers, and it wasn't even a concept. Nobody had heard of nanotechnology or synthetic biology or even nuclear weapons, right? A hundred years from now, it's likely that there might be other things that we haven't thought of.

PRACTICE

The runaway trolley returns once again in this chapter. How does the runaway trolley moral dilemma connect with concerns regarding the development of artificial intelligence and autonomous, self-driving cars? (Hint: How do you think that a very intelligent robot like "Sophia" from Hansen Robotics or Atlas from Boston Robotics would handle the two situations of the runaway trolley scenario? … now apply this to self-driving cars and describe the potential moral issue…..)



Video (4:01)

FURTHER READING

U.S. and China Compete for AI dominance May 3 2019

Purdue U. looks at 'What if AI decides to wage war?' May 14, 2019

Does Artificial Intelligence deserve the same protections we give to animals? May 9, 2019

Top Five things to know about the state of Artificial Intelligence March 3, 2020