Meeting 3: Killer Robots and AI Ethics

Philosophy of Computation at Berkeley
pocab.org

April 28, 2017

1 Impacts of Automation

Algorithms are seeping into every corner of society. Is it good or is it bad? Can algorithms solve all problems? Or are there some domains where algorithms ought not to be used? For example, should a teacher be fired based on an algorithm? Is there *any* such computable algorithm that *guarantees* some ϵ probability of correctness, and if so, can we prove that? Or can we disprove it, based on some assumptions?

> In 2007, Washington, D.C.'s new mayor, Adrian Fenty, was determined to turn around the city's underperforming schools. He had his work cut out for him: at the time, barely one out of every two high school students was surviving to graduation after ninth grade... Fenty hired an eudcation reformer named Michelle Rhee as chancellor of Washington's schools.
>
> The going theory was that the students weren't learning enought because their teachers weren't doing a good job. So in 2009, Rhee implemented a plan to weed out the low-performing teachers... Rhee developed a teacher assessment tool called IMPACT, and at the end of the 2009-10 school year the district fired all the teachers whose scores put them in the bottom 2 percent. At the end of the following year, another 5 percent...
>
> Sarah Wysocki, a fifth-grade teacher, didn't seem to have any reason to worry. She... was getting excellent review from her principal and her students' parents. One evaluation praised her attentiveness to the children; another called her "one of the best teachers I've ever come into contact with."
>
> Yet... Wysocki received a miserable score on her IMPACT evaluation... This left the district with no choie to fire her... This didn't seem to be a witch hunt or a settling of scores. Indeed, there's a logic to the school district's approach. Administrators, after all, could be friends with terrible teachers. So Washington, like many other school systems, would minimize this human bias and pay more attention to scores based on hard results: achievement scores in math and reading. The numbers would speak clearly, district officials promised. They would be more *fair*.
>
> ...Attempting to reduce human behavior, performance, and potential to algorithms is *no easy job*.
>
> ...The model itself is a black box, its contents a fiercely guarded corporate secret. This allows consultants... to charge more, but it serves another purpose as well: if the people being evaluated are kept in the dark, the thinking goes, they'll be less likely to attempt to game the system... But if the details are hidden, it's also harder to question the score or to protest against it.
>
> ...After the shock of her firing, Sarah Wysocki was out of a job for only a few days. She had plenty of people, including her principal, to vouch for her as a teacher, and she promptly landed a position at a school in an affluent district in northern Virginia. So thanks to a highly questionable model, a poor school lost a great teacher, and a rich school, which didn't fire people on the basis of their students' scores, gained one.
>
> –*Weapons of Math Destruction*, Cathy O'Neil

> [Cathy O'Neil] shares stories of people who have been deemed unworthy in some way by an algorithm. Theres the highly-regarded teacher who is fired due to a low score on a teacher assessment tool, the college student who couldnt get a minimum wage job at a grocery store due to his answers on a personality test, the people whose credit card spending limits were lowered because they shopped at certain stores. To add insult to injury, the algorithms that judged them are completely opaque and unassailable. People often have no recourse when the algorithm makes a mistake.
>
> – *Review of* Weapons of Math Destruction[1]

2 The Judgment Algoritm

Consider the following argument:

> Assume that humans are universal Turing machines, that is, a Turing machine able to execute any Turing machine whatsoever. From this, we can assume that a human H is an *arbitrary* Turing machine. Now suppose there exists a Turing machine J such that $J(H) = i$ where $i \in S$ and S is a well-ordered set of numbers. Also assume that J looks at the output of H – the output of an arbitrary Turing machine – to compute the output i. So J can be used to compare humans, such that if $J(H_1) > J(H_2)$, H_1 is more "worthy" than H_2. But H is an arbitrary Turing machine, and by the uncomputability of the halting problem, we know that J cannot know if H even halts or not! Therefore, no such J exists.

- Given the assumptions, verify that the conclusions follow, or point out how they don't.
- What assumptions were made in the above argument?
- Can the assumptions be attacked? For example, could we say that humans aren't universal Turing machines, but only capable of executing a certain set of Turing machines such that their outputs all share some property?

3 Self-Driving Cars That Kill People (And Cats)

Suppose a self-driving car has to kill person A or person B. Of course, this is not a realistic situation, because a self-driving car will rarely have a hundred percent confidence about such matters. But for the sake of argument, let's suppose that the car knows with certainty that if it take some action e_1, person A will die, and if it take some other action e_2, person B will die. There are no other available actions.

If our Judgment Algorithm J exists, the solution is simple: save A if $J(A) > J(B)$, and save B if $J(B) > J(A)$. If we can prove that J does not exist, the solution is also simple: flip a coin and kill A with $\frac{1}{2}$ probability, kill B with $\frac{1}{2}$ probability. Implicitly, this says that all people have equal moral value. [2]

But maybe this question is too hard. Maybe this question is easier: if a self-driving car has to kill either a person or a cat, which should it kill? Many of us might have the intuition that it should kill the cat. What justifies this intuition? Is it only because we are all humans? As in, if this room were full of cats rather than humans, would they all agree that the car should kill the human and not the cat? Or is this intuition more universally justified, independent of the context? Is this because we are more "complex" than cats? If so, is it moral for an AI to save a fellow superintelligent AI over a human because that AI is more "complex" than humans?

[1] https://blogs.scientificamerican.com/roots-of-unity/review-weapons-of-math-destruction/

[2] What if the car knows it can kill either person A or two people B, C? In this case, the probabilities can be multiplied, so the car kills A with probability $1 - \frac{1}{2^2}$ and B with probability $\frac{1}{2^2}$. Three people? $\frac{1}{2^3}$. And so on. This way, our intuition that a car should almost definitely kill one person over a million people is safely preserved – the chances that it kills a million people is almost nil.

4 A Moral Hierarchy of Complexity?

There is something to this idea that more "complex" beings, in both the mathematical and intuitive sense, which they mostly overlap, are more morally worthy. It forces us to draw the conclusion that more complex AI could be able to eat humans, which is spooky, but perhaps justified. Douglas Hofstadter has a few things to say about this idea:

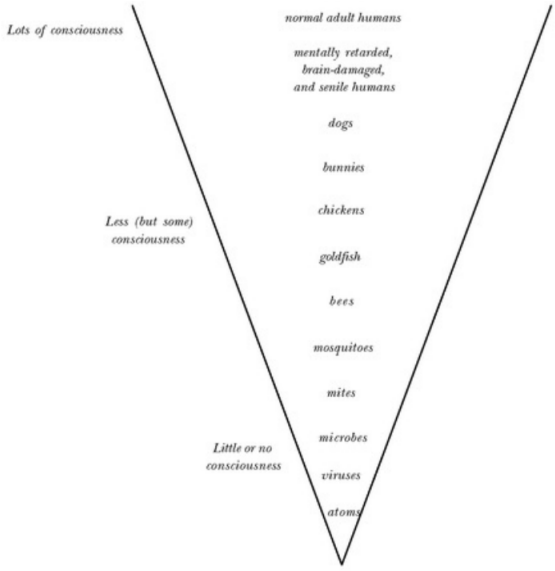
> In my teens and twenties, I played a lot of Chopin on the piano, often out of the bright yellow editions published by G. Schirmer in New York City. Each of those volumes opened with an essay penned in the early 1900s by the American critic James Huneker. Today, many people would find Hunekers prose overblown, but I did not; its unrestrained emotionality resonated with my perception of Chopins music, and I still love his style of writing and his rich metaphors. In his preface to the volume of Chopins tudes, Huneker asserts of the eleventh tude in Opus 25, in A minor (a titanic outburst often called the Winter Wind, though that was certainly neither Chopins title nor his image for it), the following striking thought: Small-souled men, no matter how agile their fingers, should not attempt it.
>
> I personally can attest to the terrifying technical difficulty of this incredible surging piece of music, having valiantly attempted to learn it when I was around sixteen and having sadly been forced to give it up in mid-stream, since playing just the first page up to speed (which I finally managed to do after several weeks of unbelievably arduous practice) made my right hand throb with pain. But the technical difficulty is, of course, not what Huneker was referring to. Quite rightly, he is saying that the piece is majestic and noble, but more controversially, he is drawing a dividing line between different levels or sizes of human souls, suggesting that some people are simply not up to playing this piece, not because of any physical limitations of their bodies, but because their souls are not large enough. (I wont bother to criticize the sexism of Hunekers words; that was par for the course in those days.)
>
> This kind of sentiment does not go down well in todays egalitarian America. It would not play in Peoria. Quite frankly, it rings terribly elitist, perhaps even repugnant, to our modern democratic ears. And yet I have to admit that I somewhat agree with Huneker, and I cant help wondering if we dont all of us implicitly believe in the validity of something vaguely like the idea of small-souled and large-souled human beings. In fact, I cant help suggesting that this is indeed the belief of almost all of us, no matter how egalitarian we publicly profess to be.
>
> ...
>
> All human beings at least all sufficiently large-souled ones have to make up their minds about such matters as the swatting of mosquitoes or flies, the setting of mousetraps, the eating of rabbits or lobsters or turkeys or pigs, perhaps even of dogs or horses, the purchase of mink stoles or ivory statues, the usage of leather suitcases or crocodile belts, even the penicillinbased attack on swarms of bacteria that have invaded their body, and on and on. The world imposes large and small moral dilemmas on us all the time at the very least, meal after meal and we are all forced to take a stand. Does a baby lamb have a soul that matters, or is the taste of lamb chops just too delicious to worry ones head over that? Does a trout that went for the bait and is now helplessly thrashing about on the end of a nylon line deserve to survive, or should it just be given one sharp thwack on the head and put out of its misery so that we can savor the indescribable and yet strangely predictable soft, flaky texture of its white muscles? Do grasshoppers and mosquitoes and even bacteria have a tiny little light on inside, no matter how dim, or is it all dark in there? (In where?) Why do I not eat dogs? Who was the pig whose bacon I am enjoying for breakfast? Which tomato is it that I am munching on? Should we chop down that magnificent elm in our front yard? And while Im at it, shall I yank out the wild blackberry bush? And all the weeds growing right by it?
>
> What gives us word-users the right to make life-and-death decisions concerning other living creatures that have no words? Why do we find ourselves in positions of such anguish (at least for some of us)? In the final analysis, it is simply because might makes right, and we humans, thanks to the intelligence afforded us by the complexity of our brains and our embeddedness in rich languages and cultures, are indeed high and mighty, relative to the lower animals (and vegetables).

By virtue of our might, we are forced to establish some sort of ranking of creatures, whether we do so as a result of long and careful personal reflections or simply go along with the compelling flow of the masses. Are cows just as comfortably killable as mosquitoes? Would you feel any less troubled by swatting a fly preening on a wall than by beheading a chicken quivering on a block? Obviously, such questions can be endlessly proliferated (note the ironic spelling of this verb), but I will not do so here.

Below, I have inserted my own personal consciousness cone. It is not meant to be exact; it is merely suggestive, but I submit that some comparable structure exists inside your head, as well as in the head of each language-endowed human being, although in most cases it is seldom if ever subjected to intense scrutiny, because it is not even explicitly formulated.

In one of the Star Wars films, I recall seeing a huge squadron of hundreds of uniformly marching robots - and when I say uniformly, I mean really uniformly, with all of them strutting in perfect synchrony, and all of them featuring identical, impassive, vacuous, mechanical facial expressions. I suspect that thanks to this unmistakable image of absolute interchangeability, virtually no viewer feels the slightest twinge of sadness when a bomb falls on the charging platoon and all of its members - these factory-made creatures - are instantly blown to smithereens. After all, in diametric opposition to C-3PO and R2-D2, these robots are not creatures at all - they are just hunks of metal!... What is it, then, that gives us the undeniable sense that C-3PO and R2-D2 have a light on inside?

– *I Am a Strange Loop*, Douglas Hofstadter



5 Human-Compatible AI and Russell's Cat

Professor Stuart Russell, an AI professor here at UC Berkeley, is very concerned about the potential negative implications of actually intelligent AI. He recently started the Center for Human-Compatible AI.

> The goal of CHAI is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems.
>
> Artificial intelligence research is concerned with the design of machines capable of intelligent behavior, i.e., behavior likely to be successful in achieving objectives. The long-term outcome of AI research seems likely to include machines that are more capable than humans across a wide range of objectives and environments. This raises a problem of control: given that the solutions

> developed by such systems are intrinsically unpredictable by humans, it may occur that some such solutions result in negative and perhaps irreversible outcomes for humans. CHAI's goal is to ensure that this eventuality cannot arise, by refocusing AI away from the capability to achieve arbitrary objectives and towards the ability to generate provably beneficial behavior. Because the meaning of beneficial depends on properties of humans, this task inevitably includes elements from the social sciences in addition to AI.
>
> –humancompatible.ai/about

In his *Times* article *Moral Philosophy Will Be Part of the Tech Industry*, Russell proposes a simple but thought-provoking question: if a housemaid robot sees that the children are hungry and the parents aren't home, what will stop the robot from cooking the house cat and serving it to the children?

> A lowly domestic robot that doesn't understand human values may do something silly – like cooking the cat for dinner when the fridge is emptyand that will be the end of the domestic robot industry. Strange as it may seem, moral philosophy will become a key industry sector. The output could be quite instructive for the human race as well as for the robots.
>
> –*Moral Philosophy Will Be Part of the Tech Industry*, Stuart Russell[3]

- What's wrong with cooking the cat? If we examine our intuition for why it is wrong to kill the cat, is it based on cultural (and therefore arguably arbitrary) values, or universal values that are preserved across cultures?
- What could stop the robot from cooking the cat? Again, is this action considered immoral because an alive cat is more complex than, for example, a piece of dead beef in the fridge, or a stalk of celery?

6 Universal Morality, Mathematical Morality

> Boredom is the root of all evil.
>
> –Søren Kierkegaard

Consider the following argument:

> Suppose there is a universal aspect to morality. By that I mean everyone, every single person in the world, or indeed any sufficiently complex being in the universe, agrees that some things are "good", that those things should happen more and ideally all the time, and some things are "bad", that those things should happen less and ideally never. If it is *universal*, then it can be *mathematicized*. Therefore, some mathematical equation for morality exists.

- Given the assumptions, verify that the conclusions follow, or point out how they don't.
- What assumptions were made in the above argument?
- Can the assumptions be attacked? For example, could we say that there is no universal aspect to morality? If we choose to take that position, what conclusions are we forced to draw? For example, if there is absolutely no universal aspect to morality, does this mean that it might be considered moral in some societies to kill each other and to kill themselves?

[3] http://time.com/collection-post/4026723/stuart-russell-will-ai-overtake-humans/