Ab uno disce omnes: Some thoughts on artificial intelligence

By Harris M

There is a marked difference between the perception and reality of Artificial Intelligence. Science fiction writers, film directors, and artists would have you believe that AI is manifested in sentient robots. This kind of AI is referred to in academic literature as “strong” AI, which is an AI that is intelligent enough to understand or learn any intellectual task that a human being can. We don't have general AI yet. The most common and widely used AI is “narrow” AI, which is intended to do a limited task, which are AI that have been trained to perform extremely specific tasks such as identifying pictures of cats or suggesting an enjoyable movie. You may have heard that Google’s DeepMind AI was able to defeat the world champion at Go, which is one of the most difficult games in the world. There are also “weak” AIs that can implement a limited part of a mind, but not the entire mind. The most common and annoying examples include voice assistants and automated chat assistants. These AIs have a very narrow spectrum they can operate in, and asking any question outside of their limited scope will result in unhelpful responses. It is important to realize that artificial intelligence is on the rise - just slower than the media coverage would lead you to believe.

Even though we don’t have general AI, there have already been a plethora of concerns that have arisen. One of the primary issues is that artificial intelligence will disrupt menial jobs. What do we do with the suddenly jobless people? This will no doubt shift the job demand towards fields such as software engineering and robotics, but there will be notable disruptions in areas such as manufacturing and data entry. In short, any task that is simple enough to be learned by a computer will more than likely be taken over by a computer. This is widely viewed as a good thing, allowing humans to focus on their highest value tasks, saving huge amounts of time, which can then be used for analyzing, optimizing, and ultimately making better decisions. Furthermore, how do we redistribute the wealth generated by machines? Companies such as Alibaba and Amazon now have warehouses entirely staffed by robots. Should the money made by these warehouses go to the company? The workers? The government? This is a difficult question to answer, but should absolutely be addressed if we are to move forward.

Even though we don't have general AI, there have already been a slew of ethical and practical concerns that have arisen. In the following paragraphs, we will briefly explore a few of these concerns.

One of the main issues with AIs is that they are not perfect. AIs are trained for extremely high accuracy (usually, anything below 90% is considered unusable)) and are constantly improving. There are no guarantees that the AI will not result in disastrous results in high-risk applications such as self-driving cars or the medical field. Notably, Tesla’s self-driving AI has proved to be fatal for a driver when it crashed into a white semi-truck when it could not distinguish it against a bluish-white sky. Too often, Artificial Intelligence is asked to do herculean tasks which often yield devastating results. Consider IBM’s “Watson for Oncology” AI. This was an AI that would be programmed to recognize and recommend cancer treatment guidelines to biomedical researchers. This was meant to aid cancer patients by suggesting unique treatments. The final implementation was deemed by healthcare professionals to be an utter failure, as Watson suggested incorrect and exceedingly dangerous cancer treatment advice. Microsoft’s Twitter chatbot was meant to converse with twitter users, and it only took one day for the internet to teach it to be misogynistic, racist, and anti-semitic. This is a good example that it can be

extremely hard (almost impossible) to anticipate all the ways in which AI can be misused. Bias is a huge issue in AI; it is often said that a programmer will write his/her bias into every program they write, and AI is no exception. Amazon's recruitment AI had to be deprecated because it would disproportionately choose males over females. This example actually points to a larger issue: bad datasets. Bad datasets can lead to devastating misclassifications. An AI is only as good as the data used to train it, and the Amazon AI was trained with current engineering employees, which is a predominantly male field. Google Photos image classification algorithm was probably the most egregious example of AIs being wrong.: it classified African American people as Gorillas. Their solution? Remove the Gorilla label from the classification corpus. This is clearly not the correct way to go about this, but showcases that AI still has a long way to go, and we are not entirely sure how to deal with failures. There have already been a plethora of examples of artificial intelligence being used wrong, and it is important that we prevent this from ever happening again. These are not insurmountable challenges, but they do highlight that AI should not be allowed to run amok.

Up until now, we have only been discussing the issues with narrow and weak AIs, but inevitably, general AI will be developed and have similar issues. There will be a massive surge in AI when it is able to program itself to do anything. If AIs were sentient (very likely, as we seek to emulate humans), how would we control them? One of the primary reasons that artificial intelligence is hard is that it is hard to control. How do we ensure that these sentient AI will not decide to fight back? There will undoubtedly be discussion of giving robots rights, which is a much more complex matter altogether, and none of these have any clear answers.

Human-robot interaction is a field that has been rapidly growing without much regulation. Notably, sex robots have been sold that look and sound exactly how the user wants. The implications of this are terrifying: customers have already begun ordering uncomfortably young-looking robots and are recreating rape scenarios, which is horrifying. There are obvious psychological effects that go along with this, but these have not been explored in any fashion. The film Blade Runner 2049 shows an artificial intelligence projected as a hologram that is intended to be a fully customizable romantic companion. Similar to sex robots, this raises the questions of what kind of ramifications this will have on people and ultimately on the human race. There was a marriage between a man in Japan and an artificial intelligence with no physical form. The most terrifying (and awesome) form of human-robot interaction is brain-computer interfaces(BCIs). Initial development is aimed at aiding people with neurological disorders, but it will not take long for this to extend to cybernetic modifications on people with no disorders. It is hard to anticipate how far AIs can be used.

Similar to BCIs, deepfakes are simultaneously terrifying and awesome. Deepfakes are AIs that can be programmed to look and sound exactly how the user wants. The problem with deepfakes is that they often lead to imitations that are indistinguishable from the original. Deepfakes began circulating in 2017, which were humorous at first, but the danger of deepfakes quickly became apparent when (unsurprisingly, given the context) it went towards pornography. According to a study, a staggering 96% of deepfakes are pornographic, and started as deepfaking celebrities into pornographic films but quickly spread to revenge porn, fabricating clips of people saying things they didn't, and financial hoaxes. The widespread exploitation of deepfakes led to DARPA funding a project to detect deepfakes on the internet, and academic research has focused around detecting deepfakes. Some platforms, such as Facebook,

instead have opted to enforce that users clearly note that media is a deepfakes. Some AIs are much more nuanced, and thus that much more dangerous. GPT-3, a text generation AI developed by OpenAI, is able to generate text so similar to that written by humans that they did not release it to the public out of fear that it would be rampantly misused. OpenAI has also released DALLE, a model that can generate artwork based on text prompts, which is a fun experiment, but could easily be misused.

One of the biggest issues with AI is that it is incredibly hard to relay these ideas to people outside academia, which is what leads to the difference between the perception and reality of AI. AI developers should be straightforward and honest about what is being developed, and there needs to be an effort to educate on what AI can do. AI tends to be a very mysterious field to those not in, but if the basics were taught, there would be a lot less misunderstanding about what AI can and cannot do. The problem is that most people are unfamiliar with the psychology involved, the cultural implications, and the practical issues of developing AIs. It is important to realize that artificial intelligence is not a kind of magic; rather, it is a science that should be applied with the highest level of education possible.

In the spirit of clarity, I must admit that Harris did not write this essay. This essay about AI was written by an AI that has read several samples of Harris's writing and then answered a series of prompts to generate this essay, best attempting to mimic his style. This essay is designed to draw the reader's attention to the fact that there are significant problems with the way that AI is being used and developed. It is my hope that this essay serves as a useful reference point when it comes to how to think about the implications of AI, and ultimately, help to create safer, more equitable, and more than a little terrifying environments for AI to operate in. The point is that we don't have a crystal ball and it is entirely possible that we could create a machine consciousness far beyond what is currently possible. This could have dire consequences for human-robot interaction, health care, and ultimately the human race. I strongly suggest learning about what AI can and cannot do, to best further humans as a species. If you don't, it is entirely possible that when general AI comes to realization, that will be the end of humanity as you know it.