Authors
Tobias FIlmberg
Mirco Ghadri
Sameer Jathavedan

Ethical AI essay

With the release of Open-AI´s Chat GPT-4 it gets clearer and clearer for the public what power resides in these technologies. The potential is huge. Tasks that previously seemed impossible and only for a human, can now be done by AI, faster and more efficiently. However, just like any invention, AI comes with its risk and a huge such risk. This essay will focus on some of the risks and problems associated with AI and potential solutions.

In the text *What worries me about AI* François Chollet talks about how he sees AI and the potential risks. He starts to compare AI to past technologies and how we always have failed to comprehend or foresee the potential risks or malicious uses of technologies. In his own words he is worried about: “the highly effective, highly scalable manipulation of human behavior that AI enables, and its malicious use by corporations and governments.“ So how do we prevent malicious use of AI? Do we now exist in a world where there is another destructive tool that to some degree is uncontrollable? Maybe.

Corporations do need to follow legislation, so there is some degree of control there. But in order to implement effective legislation, the policymakers need to be aware and educated about the problems. In America where a very large amount of tech companies reside, there seems to be a knowledge gap between legislators and the current wave of technology. If one watches the congressional hearings with for example Mark Zuckerberg, it is clear that some of the politicians in that room have no idea of the scope of the problem. And as long as that is the case, the companies will be allowed to roam free and chase profit.

Because they are chasing profit. A couple of years ago Netflix released the documentary *The social dilemma* that showcased how for example Facebook operates. Most people probably have not thought about what makes Facebook tick and this documentary really showcased the driving factors behind Facebook. It is easy to think that Facebook only operates to spread their amazing tool to the world. To enable people to connect. Everybody understands that they need to make money somehow and we have all seen their ads. But I think very few have thought about what actually is being sold. One assumes that it is ad watch-time and to some degree personalized ads. But the documentary shows that in reality it is not the ads itself that is the thing Facebook sells, it is the data they know about you.

Since they have so many of your interactions they know your interest, values and beliefs. One of the key factors to address this is, according to the article, to enable

the users to decide what and how much help they want by AI services. We think that this is a good step, but we also think that it overestimates how much the public cares or has energy to care. There are so many things in people's life that they on some level know are harmful or not beneficial, but they do not act upon. Even given choices, people tend to choose the preset or the easiest option. We think that if given the option, only a part of the population would take the time and energy to tune how the AI in the services they use operate. And if that is the case, there would still be a large portion of the people that are susceptible to information attacks or bubbles.

In the article “Who should stop unethical A.I?”, Mathew Hutson describes the importance of peer reviewing research work in Computer Science. One of the issues that is brought up is how ethics is not one of the core focuses in the peer review process, at least not for computer science. The reason is that fields such as biology and psychology deal directly with human subjects while a research paper on an AI algorithm for example, does not do that in the same way. I quote: *“University research that involves human subjects is typically scrutinized by an I.R.B.(Institutional Review Board), but most computer science doesn’t rely on people in the same way.”* However, this has started to change as researchers realize the importance of the ethical aspects of research and long term consequences. Katie Shilton, an information scientist at the University of Maryland, brings up 4 ethical aspects of AI that are potentially dangerous and need to be regulated:

1. AI that can be used as weapons against the population, such as facial recognition that can be used for mass surveillance
2. AI that can be biased towards their users such as speech to face models that perform poorly on certain people
3. AI that can be used to create dangerous weapons
4. AI that can be used to create fake images, videos or news articles

All four categories seem very relevant. For example, the creators of ChatGPT are aware that it can be used to create fake news, so they have censored that functionality from the application. This is not very different from how existing so called deepfake models create fake videos and speech. A combination of text synthesis models such as ChatGPT and deepfake models could lead to a change in how videos are created all together. How can we then trust what we see, read and hear? When it comes to AI that can be used against the population such as in mass surveillance, we believe it can be both good and bad. If used by a totalitarian government, it could be used as a tool of persecution of the population and oppressive control. This could be said about the mass surveillance system in China, where the laws of the country are questionable and ethnic minorities are being oppressed. However, surveillance systems have also been used with great success in democratic countries such as England where they help to solve criminal cases every day. So we believe that it depends on which government is using the system and for what purpose.

The solutions according to the article, lies not only in a more rigorous ethical peer review process, but also in auditing private companies that use these technologies. A problem is that the technologies are developing faster than people have time to learn, so regulatory agencies many times do not know the technologies they are auditing and how they can be misused.

These problems are non-trivial to solve. The major issue is more than the fact that regulatory organizations do not understand how AI works - it is that most of the AI technologies implemented today are fundamentally black box technologies. As Clive Thompson mentioned in the article "*Sure, A.I. Is Powerful—But Can We Make It Accountable?"*, while AI algorithms work similar to how the human brain works, we cannot scrutinize the predictions made by AI in a way that we do to humans. If we ask a bank officer why our bank loan request was rejected, they could tell us the reason - our neighborhood being a high-risk area or our credit score being low. This level of transparency is not available for current AI technology, especially deep neural networks. This makes accountability difficult in case of AI prediction methods.

Research is underway to figure out methods to probe into the workings of the black box nature of deep learning models. As mentioned in the article above, private companies such as Clarifai are working on projects to analyze what the inner layers of convolutional neural networks are seeing. Google is also developing methods to figure out hallucinogenic "deep dreaming" techniques to figure out how neural networks generate patterns. These projects give hope in the direction of bringing more transparency to the inner workings of AI algorithms.

Even so, there is sentiment in the academic sector that supports the black box nature of AI. According to the article *Can We Open the black box of AI?*, computer scientists argue that the creation of transparent AI should not be considered as a replacement for deep learning, but as a complementary effort, despite concerns about its impact. They believe that transparent techniques could be useful for problems that are already described as a set of abstract facts but may not be as effective in perception, which involves extracting facts from raw data. These scientists suggest that machine learning's complex answers should be a part of science's toolkit because the real world is intricate and complex, and reductionist, synthetic descriptions might not exist for phenomena such as the weather or the stock market. Stéphane Mallat, an applied mathematician at the École Polytechnique in Paris, emphasizes that there are certain things that cannot be verbalized.

In conclusion, AI is a complicated tool with many benefits as well as risks. Better awareness on these tools among law-makers and politicians is a necessity for a sustainable development in the field. But also a stronger effort from scientists and researchers in applying ethical scrutiny when peer reviewing each others text. Still, some areas of AI remain a mystery, such as the black box of how a deep neural network makes its decisions. But the future for AI looks promising to say the least.

References:

1. https://medium.com/@francois.chollet/what-worries-me-about-ai-ed9df072b704
2. https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai
3. https://www.wired.com/2016/10/understanding-artificial-intelligence-decisions
4. https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731