Intelligent Agents
Assignment 2.2 Ethics

Erik Norlin

March, 2023

1. The importance of ethics in AI

If you have not been living under a rock you are probably aware of that artificial intelligence has been trending a lot lately, especially with the up-rise of the infamous statistical language model (SLM) ChatGPT, as well as other SLMs like Sydney, which is an SLM combined with a search engine. Apart from SLMs, AI has made significant advancements in computer vision and classification tasks, which for example allows for highly accurate facial recognition, and shown to be incredibly powerful in these categories. This kind of AI technology is based on black box models that will get better and better and are something we are likely to see much more of in the not so far future. This is because the science and technology behind AI is growing very fast, and have been developed to the point that the technology has become good enough for normal society to find value and use of it. However, in the hectic rise of AI technology there are emerging questions and dilemmas around AI that that we as a society have never encountered before, which needs to be empathized. There are also ways where AI can be used unethically and have been used unethically which have to be brought to the light of discussion. Also, research shows that ethical guidelines for developing AI have practically zero effect on decision making of software developers [7]. This should be alarming because lack ethical consideration in combination of powerful technology can potentially lead us to face a dark future.

In the short time period that AI has penetrated mainstream society there has already emerged dilemmas about usage of AI. Specifically, black box based SLMs, like ChatGPT, has been both celebrated and condemned, celebrated particularly by people who find use of auto generating text, like students, and thus condemned by some teachers and professors. This gives rise for ethical dilemmas especially for students because auto generating text for assignments might give them better grades, but at the same time doing this is not only cheating but also defeats the purpose of learning and hence stunting growth of deeper understanding of the world. Also that those who use models like ChatGPT might have an advantage over those who do not use it, which creates an unfair competing environment between students, since some students compete to get into particular programs or courses. Teachers and professors therefore have responsibility to adjust to this new wave of technology and come forward of what are appropriate ways of examining students that make sure that students learn and can prove their own understanding and also offers fair competing environments for the students, because black box based SLMs are here to stay and will continue to advance.

Another ethical dilemma with black box based SLMs, that perfectly demonstrates the importance of ethical consideration when developing AI algorithms, is that they do not inherit common sense but at the same time are often designed to be agreeable towards the user. If a user says to ChatGPT ”I feel dirty, should I take a shower?” ChatGPT could respond with saying ”Yes, I think you should”. This mechanism of agreeableness could however be a recipe for a catastrophe. In one study where researchers were using OpenAI’s GPT-3 as a conversational agent to simulate a person having depression, it advised the fake patient to kill themselves [5].

Even though the technology of black boxes has shown to be impressive with high capability, the technology could also be used to impose harm. With great advancements within computer vision and classification where AI performs phenomenally well at facial recognition there lies a real danger that we could perhaps head into an Orwellian dystopia where companies and/or governments would be the Big Brother watching us. However, many of us today are already being ”watched” because with the use of computers, smart phones etc. companies are constantly gathering data about us through these devices and not many people seem to bother with it. Although this is the case, with the advancement of AI within computer vision, the thought of being surveilled by authorities and knowing exactly where you are and what you are doing at any time is very uncomfortable to think about, especially if authorities are, or would become, corrupt. It is therefore critical to emphasize the importance of ethics in AI and discuss how to align ethics with this fast evolving technology, because it is not obvious how to do this once one starts digging into potential problems like these.

Another ethical dilemma that arises with the advancement of AI are autonomous driving vehicles that could perhaps one day replace human drivers all together. If an autonomous driving vehicle ends up in a situation where it is faced with either driving over a child or driving into a wall and killing the passenger inside the vehicle, what should the vehicle do? Who is responsible? The person inside the self-driving vehicle? The company that made the vehicle? The vehicle itself? Dilemmas like these start getting complex fast as one can see here. There are numerous of other examples of dilemmas that are not covered here but these should make it clear enough how important it is to consider ethics when using and developing AI.

2. Examples of unethical uses of AI

With the potential that AI has, there are infinite examples of how the advancement of AI could lead to unethical usage. There are some especially concerning examples however, that actually have taken place.

On November 15th last year Meta released a new large black box based SLM called ”Galactica” that was intended to be an assisting tool for the science communities, and was trained on scientific material across the internet. However, Galactica was taken down just a few days after it had been released due to heavy criticism. Apparently, Galactica had been creating new scientific papers about completely made up information and theories [8], [1]. For instance, someone reported Galactica producing a paper about benefits of eating crushed glass. This is obviously not true and anyone can figure that out. The problem is when a model like this starts producing papers that are not as easy to tell if they are true or not and those papers becoming available online. This could be the start of a huge snowball effect of spread of misinformation in science, which if anything has to be as factually correct as we can make it. If we could not distinguish real papers from fake ones, this would be disastrous indeed. Galactica was naively put online in hope to advance science, which was unethical in the sense that naivety should not be an excuse for disastrous, dangerous outcomes. Furthermore, an AI powered SLM that creates false information and makes it available online could also be used unethically with the purpose of steering or swaying people into believing in falsehoods, whether it is within science, politics, media or anything else.

With powerful AI driven computer vision, deep fakes of faces and voices could be unethically used to cause harm to other people’s personas, as well as being used to scam people. According to some officials, the first deep fake scam they ever heard of occurred in 2019, when the CEO of a UK-based energy company was scammed of $243,000 when his "boss" called him telling him to make an urgent transaction to a supplier. It turned out that the caller was someone using an AI algorithm to replicate the voice of his boss [15]. Furthermore, deep fakes can also be unethically used to sway public opinion about politicians and even increase political polarization. In the beginning of the war between Russia and Ukraine, Russia released a video of a deep fake video of Ukrainian President Volodymyr Zelensky telling the Ukrainian army to stand down. Luckily, social media companies acted quickly and took down the video [3]. Since AI is getting better and better, deep fakes will get better too. This means that we are going to have to be even more skeptical of what we see and listen to online.

One enormous issue with black box based AI is its' tendency to be biased against groups of minorities in classification tasks because AI gets as good as the data it is trained on, which can be incomplete or contain biased content. For example, clinicians may incorrectly reject diagnosis of heart attacks in older women because this group category are more likely to have unusual symptoms. An AI algorithm may therefore recommend clinicians to test for something else other than a heart attack, delaying potentially life saving treatment [13]. There are plenty examples of people that have been treated incorrectly in high stake decision making due to these kinds of biases in AI systems. Partly what is unethical about this is that companies develop these kind of AI algorithms and sell them for enormous amounts of money and do not reveal these kind of flaws in their models because first, these models are often propriety. Second, saying that their products are biased against minorities is not a very efficient marketing strategy.

Another thing regarding ethics in AI with the advancement of it is the fact that the progress of AI technology could, if it has not already happened, spark an AI race between world powers, i.e. spark politics of insecurity. Meaning that the world powers would be racing against each other to have the best AI military technology in fear of being inferior and run over by another nation with superior AI military technology if the nation does not keep the pace in the race. Implications of this could lead to cutting corners of safety standards in the advancement of AI and increase risk of real conflict, i.e. compromising ethics in usage and development of AI [7], [2].

3. Case study: Mass Surveillance

Mass surveillance using AI technology is arguably one of the greatest dangers and unethical use of AI that the world is facing currently that could happen in the near future, if it already has not starting to happen. As mentioned before, AI technology performs phenomenally well when integrated with computer vision which allows for highly accurate facial recognition. This opens up better possibilities for identification of people for better or worse. Technology itself is never evil, but in the wrong hands there is always a great risk of unethical usage of it. Mass surveillance of people in some regions of the world are no news, China has since 2005 integrated a mass surveillance system called "SkyNet", which is unironically the same name as the computer system that was out to destroy mankind in the Terminator movies [11]. In 2015 China had covered 100 percent of Beijing with SkyNet, really showing how serious they take surveillance and social control [4]. China's reason for implementing mass surveillance is to hunt down fugitive corrupt officials and criminals [9], which can be considered reasonable, but implications of mass surveillance also comes with a price for the people. The regime of China has in the last few years intensified surveillance on minorities in the country, mostly on Muslim Uyghurs who the Chinese government consider their values and beliefs threatening to the regime of the country [12]. Furthermore, the Chinese Communist Party's (CCP) law enforcement agencies continuously harass and violate human rights of people that have opposing values and beliefs that they consider threatening to the regime. CCP's biggest target group that they recognize as most threatening are Turkic-speaking Muslim Uyghurs, as well as writers and activists [16]. Some human rights groups have claimed that the regime of China has kept more than a million Muslim Uyghurs against their will in ”re-education” camps, where crimes against humanity has been reported [17]. The intensification of surveillance on minorities are not only consisting of cameras but also of monitoring of peoples behaviour such as electricity use, how often they use their front door etc [17]. Similarly, the regime of China has plans on implementing a mandatory ”social credit score” for everyone that would rank their social behaviour. For example, your social credit score would go down if you would drive badly or buy too many video games [10]. One could argue that China is already living under an Orwellian dystopia where ”Big Brother is watching you”, where some might feel safer and some definitely not.

With the incredible advances in AI in the last couple of years, the regime of China could potentially use AI as ”steroids” for their mass surveillance, as if the current situation is not worrying enough. In a not so far future, AI facial recognition could potentially identify almost every action you take and this would directly impact your social credit score. AI facial recognition could also potentially be abused to classify facial features of minorities or specific people and from there hinder these people in public for example. AI powered mass surveillance could also be more than just facial recognition, it could just as well be used for classifying what you are allowed to do and not do in society based on information about your behaviour that the state constantly gathers from you, not only from cameras. With tracked and stored information about your behaviour AI could predict your choices and could therefore allow or deny permission of what you can do and not do in any given situation. For example, if you can take loans, if you can travel out of the country, where you can go publicly, who you can meet, and lots of other things. This could arguably very well be the deepest abyss of unethical usage of AI in society as one can imagine, that an Orwellian dystopia would become reality with the help of AI.

4. Preventing unethical uses of AI

To move in a direction towards a future of ethical awareness when using and developing AI in the explosion of its’ upcoming, it is reasonable to assume that change has to start happening in the beginning of the supply chain, where software developers are sculpturing the future technology of AI. As previously mentioned, ethical guidelines have practically zero impact on software developers within development of AI, but according to some studies there are numerous ways of empathizing ethics in AI to start making this change.

In order to tackle this challenge, Strümke, Slavkovik and Madai (2022) proposes that AI development should be its’ own profession, like medicine for instance, and that a professional AI developer should have a license, as a doctor in medicine, to be able to work within the field. This license would inherit a code of conduct of ethics in AI one must follow when developing AI technology, and without this license one could not work in the field of AI development. This would mean that if an AI developer is part of unethical development or usage of AI the person could lose their license [14]. This could make it more difficult for companies and states to develop AI technology with unethical intentions or without any ethical intentions because AI developers would be more motivated to engage ethically to avoid losing their licenses. This kind of rule obedience can be categorised as a so called ”deontological” ethical approach that is an approach of following strict rules. This is the most common way of approaching AI ethics, which perhaps makes sense because strict rules concretize what should and what should not be done in a corporate environment. There are however another approach called ”virtue” ethics which is an approach of listening and acting according to ones inner moral compass. Virtue ethics are more vague but to encourage this kind of ethics in a corporate environment could perhaps have an impact on everyday decision making in the development of AI. Empathizing virtue ethics would imply that everyone involved in the development of AI would have to take responsibility for the implications of their action [7]. This approach however could potentially be difficult to see too much of an impact from. It is arguably naive to think that encouragement of virtue ethics would cause an ethical shift to happen in the corporate world, but it is still an approach worth considering.

Regarding ethical guidelines of AI as previously mentioned, one way of making these more effective is to create technical instructions of how to develop AI ethically. The reason for this is that developing AI with ethics in consideration can be difficult if the values and principals of the ethical guidelines are highly abstract. For instance, what does a "human-centered" AI look like? Making ethical guidelines more technical, i.e. microethical, would make it a lot more convenient to implement. Partly the reason why ethical guidelines are too abstract for software developers to make effort considering them is that there is a gap of understanding between software developers and ethicists, where ethicists are lacking technical understanding of AI technology. Ethicists and software developers need to come together and discuss what ethical implementation actually means on a more practical level. The alternative is to play the blame game and say that the other party is responsible for the lack of ethical usage and development of AI. Shifting abstract principles to microethics can lead to influencing software developers to make more ethical everyday-choices in the development of AI. An example of microethics is to introduce standardized datasheets that contain descriptions of properties of data sets, such as the purpose of the data set, the best use for it, how it was created etc. This could lead to software developers of AI making more informed and ethical choices in development of the technology [7], [6].

5. Conclusion

AI will continue to advance and be integrated more and more into society, probably very quickly considering how fast the science of it progresses. We have to be aware of arising ethical dilemmas that come with the advancement of AI. Auto generating texts, chatbots that lack common sense, autonomous vehicles, deep fakes are a few of many examples where ethical dilemmas occur as we have investigated. Also concerning is unethical use of AI that has been observed, and could potentially happen. The release of Galactica, biases against minorities, AI race between world powers, money scams and political swaying using deep fakes are all real examples of this in the short time period that AI has progressed significantly. With the power of AI it is a real possibility that corrupt states can leverage AI to gain social control, potentially allowing an Orwellian dystopia to become a reality where "Big Brother is watching". In the development of AI, research shows that ethical guidelines of AI practically has zero impact on software developers, which is concerning indeed. However, despite dilemmas and dangers that come along with AI there are numerous things we can do to proceed towards a safer future where AI is used and developed with more ethical consideration. Making "AI developer" its' own profession that holds a license with ethical codes of conducts could motivate software developers of AI to make more ethical decisions in their work. Encouraging virtue ethics regarding AI in work places could be another way of influencing software developers to make more ethical decisions. Ethical guidelines in AI are too abstract for software developers to have any effect and need to be concretized and become more technical. This would make it more straight forward for software developers of AI to make informed and ethical choices in the process of development. There are certainly more to explore on the topic of ethical use in AI and it is important that we start taking ethics in AI more seriously. On an ending note, one day AI might be a part of everyone's life and it is therefore important that we all take on some responsibility regarding usage, development and implementation of AI, because everyone is a part of building the world of future.

References

[1] Mattias Bastian. *Danger to science: researchers sharply criticize Meta's "Galactica"*. Nov. 2022. URL: https://the-decoder.com/danger-to-science-researchers-sharply-criticize-metas-galactica/ (visited on 02/2023).

[2] Stephen Cave and Seán S ÓhÉigeartaigh. "An AI race for strategic advantage: rhetoric and risks". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 36–40. Cited by 98.

[3] Jack Cook. *Deepfake Technology: Assessing Security Risk*. July 2022. URL: https://www.american.edu/sis/centers/security-technology/deepfake_technology_assessing_security_risk.cfm#:~:text=Often%5C%2C%5C%20they%5C%20inflict%5C%20psychological%5C%20harm,technology%5C%20to%5C%20conduct%5C%20online%5C%20fraud. (visited on 02/2023).

[4] C. Custor. *Skynet achieved: Beijing is 100% covered by surveillance cameras, and nobody noticed.* Oct. 2015. URL: https://www.techinasia.com/skynet-achieved-beijing-100-covered-surveillance-cameras-noticed (visited on 02/2023).

[5] Ryan Daws. *Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves.* Oct. 2020. URL: https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/ (visited on 02/2023).

[6] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92. Cited by 1150.

[7] Thilo Hagendorff. "The ethics of AI ethics: An evaluation of guidelines". In: *Minds and machines* 30.1 (2020), pp. 99–120. Cited by 823.

[8] Will Douglas Heaven. *Why Meta's latest large language model survived only three days online.* Nov. 2022. URL: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (visited on 02/2023).

[9] Keira Lu Huang. *China is using an operation called 'Skynet' to track down fugitive corrupt officials.* Mar. 2015. URL: https://www.businessinsider.com/china-is-using-an-operation-called-skynet-to-track-down-fugitive-corrupt-officials-2015-3?r=US&IR=T (visited on 02/2023).

[10] Aaron Mok & Katie Canales. *China's 'social credit' system ranks citizens and punishes them with throttled internet speeds and flight bans if the Communist Party deems them untrustworthy.* Nov. 2022. URL: https://www.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4?r=US&IR=T (visited on 02/2023).

[11] Frank Langfitt. *In China, Beware: A Camera May Be Watching You.* Jan. 2013. URL: https://www.npr.org/2013/01/29/170469038/in-china-beware-a-camera-may-be-watching-you (visited on 02/2023).

[12] Lindsay Maizland. *China's Repression of Uyghurs in Xinjiang.* Sept. 2022. URL: https://www.cfr.org/backgrounder/china-xinjiang-uyghurs-muslims-repression-genocide-human-rights (visited on 02/2023).

[13] Ravi B Parikh, Stephanie Teeple, and Amol S Navathe. "Addressing bias in artificial intelligence in health care". In: *Jama* 322.24 (2019), pp. 2377–2378. Cited by 256.

[14] Inga Strümke, Marija Slavkovik, and Vince Istvan Madai. "The social dilemma in artificial intelligence development and why we have to solve it". In: *AI and Ethics* 2.4 (2022), pp. 655–665. Cited by 7.

[15] Catherine Stupp. *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case.* Aug. 2019. URL: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (visited on 02/2023).

[16] Jane Tang. *China casts its 'SkyNet' far and wide, pursuing tens of thousands who flee overseas.* May 2022. URL: https://www.rfa.org/english/news/china/skynet-repatriation-05042022151054.html (visited on 02/2023).

[17] *Who are the Uyghurs and why is China being accused of genocide?* May 2022. URL: https://www.bbc.com/news/world-asia-china-22278037 (visited on 02/2023).