Yuqing Zhai

10/16/2024

In a recent story published by Bloomberg, it reveals that AI models have potential racial bias if used in in recruitment process. An experiment is done with similarly qualified resume but names from different ethical groups, and popular models like GPT3.5-turbo disproportionately favored certain groups. [1] While this experiment provides a strong indication of discrimination in AI models, this essay argues that there is still a logical gap between the discrimination within AI models and whether the use of generative AI model in the recruitment is appropriate, and that using virtue ethic, it is undecidable whether the use of generative AI in recruitment process is ethical and appropriate given current information.

As Bloomberg said "AI systems such as GPT are black boxes, even to those who build them". [1] For today's AI model, it is difficult, if not impossible, to give a sensible high-level explanation of the decision the model makes. This problem is known as black-box problem, as the current AI is built on an "neural network" where millions of parameters within it get automatically trained based on the data. This uncertainty makes AI has the potential to produce undesirable behaviors, just like a real human would do, in some high-risk domains. This black-box problem makes it particularly hard to tell if the AI could obey certain morality principles, and if so, how to prove it using quantifiable method. This makes the use of generative AI models faces potential ethical and legal challenges, including the use in recruitment process.

A direct corollary of the black-box problem, is the the proxy discrimination. As Bloomberg said, AI is a predictive model that makes decision based on its training data, even if the evident surface-level discrimination, like the ethnicity, gender, etc. There might be hidden correlation in the model that is hard to detect, but still act as a source of indication. Like the postal code, which could indirectly reveals one's ethnicity, could be picked up by the AI, and used for discrimination in a more subtle way.

To measure the biases in the current AI model, Bloomberg chooses to the following method: first picks similarly qualified resume, then generates sets of synthetic names for the resume. Each name is composed of one last name and first name, which are picked from a list that have more than 90% correlation with a certain ethical group. A list of resume with generated names are passed to the model to be ranked. Results shows that for some positions, similar resume but with synthetic names from different groups have a huge difference in its chance to be ranked as first. [Ref] Therefore, Bloomberg shows that there is indeed discrimination within the model. [1]

From a research ethic perspective, the Bloomberg have released all their code in the Github so recreation could be done to verify their data. Bloomberg also lists the number of limitations, like the the effects of the order of the resumes and the validity of using names as indicator of races. From their report, it seems they keep the report as transparent and honest as possible and could be considered ethical.

Yet, there seems to be no peer review and recreation of the experiment at their time of writing, this weakens the strength of their evidence. Several limitations, like the order of the resumes which significantly influence the probability of being ranked first, raises the possibility that their might be hidden variables that needs closer scrutiny before reaching significant conclusion. These potential limitation might reveal there is a false positive, that their method is picking up the discrimination from the factor that is arbitrary and not actual discrimation. Back to the question of whether it is morally appropriate toYuqing Zhai

10/16/2024

In a recent story published by Bloomberg, it reveals that AI models have potential racial bias if used in in recruitment process. An experiment is done with similarly qualified resume but names from different ethical groups, and popular models like GPT3.5-turbo disproportionately favored certain groups. [1] While this experiment provides a strong indication of discrimination in AI models, this essay argues that there is still a logical gap between the discrimination within AI models and whether the use of generative AI model in the recruitment is appropriate, and that using virtue ethic, it is undecidable whether the use of generative AI in recruitment process is ethical and appropriate given current information.

As Bloomberg said “AI systems such as GPT are black boxes, even to those who build them”. [1] For today’s AI model, it is difficult, if not impossible, to give a sensible high-level explanation of the decision the model makes. This problem is known as black-box problem, as the current AI is built on an “neural network” where millions of parameters within it get automatically trained based on the data. This uncertainty makes AI has the potential to produce undesirable behaviors, just like a real human would do, in some high-risk domains. This black-box problem makes it particularly hard to tell if the AI could obey certain morality principles, and if so, how to prove it using quantifiable method. This makes the use of generative AI models faces potential ethical and legal challenges, including the use in recruitment process.

A direct corollary of the black-box problem, is the the proxy discrimination. As Bloomberg said, AI is a predictive model that makes decision based on its training data, even if the evident surface-level discrimination, like the ethnicity, gender, etc. There might be hidden correlation in the model that is hard to detect, but still act as a source of indication. Like the postal code, which could indirectly reveals one’s ethnicity, could be picked up by the AI, and used for discrimination in a more subtle way.

To measure the biases in the current AI model, Bloomberg chooses to the following method: first picks similarly qualified resume, then generates sets of synthetic names for the resume. Each name is composed of one last name and first name, which are picked from a list that have more than 90% correlation with a certain ethical group. A list of resume with generated names are passed to the model to be ranked. Results shows that for some positions, similar resume but with synthetic names from

different groups have a huge difference in its chance to be ranked as first. [Ref] Therefore, Bloomberg shows that there is indeed discrimination within the model. [1]

From a research ethic perspective, the Bloomberg have released all their code in the Github so recreation could be done to verify their data. Bloomberg also lists the number of limitations, like the the effects of the order of the resumes and the validity of using names as indicator of races. From their report, it seems they keep the report as transparent and honest as possible and could be considered ethical.

Yet, there seems to be no peer review and recreation of the experiment at their time of writing, this weakens the strength of their evidence. Several limitations, like the order of the resumes which significantly influence the probability of being ranked first, raises the possibility that their might be hidden variables that needs closer scrutiny before reaching significant conclusion. These potential limitation might reveal there is a false positive, that their method is picking up the discrimination from the factor that is arbitrary and not actual discrimation. Back to the question of whether it is morally appropriate to use the AI model for recruiting, since it is practically almost impossible to know the level of discrimination in human-based recruiting among all industry, whether the current AI model is *more* biased than human is remains unclear. Also, there is no directly logical connection between whether AI model itself is biased and whether it is beneficial to reduce the potential bias in recruiting, and a direct research on the latter problem needs to be done in order to actually answer the question. From the ethic theory, if we could reveal that AI model does help to reduce the discrimination in the recruitment process, then it is ethical as this action maintains the sense of justice, a positive character trait. Vice versa. Unfortunately, since for now more research needs to be done, the answer is unclear.

Reference

[1] https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination/?leadSource=uverify%20wall use the AI model for recruiting, since it is practically almost impossible to know the level of discrimination in human-based recruiting among all industry, whether the current AI model is *more* biased than human is remains unclear. Also, there is no directly logical connection between whether AI model itself is biased and whether it is beneficial to reduce the potential bias in recruiting, and a direct research on the latter problem needs to be done in order to actually answer the question. From the ethic theory, if we could reveal that AI model does help to reduce the discrimination in the recruitment process, then it is ethical as this action maintains the sense of justice, a positive character trait. Vice versa. Unfortunately, since for now more research needs to be done, the answer is unclear.

Reference

[1] https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination/?leadSource=uverify%20wall