The Injustice of Artificial Intelligence: Impartiality with Bias

Chanut Naktranun* and Sivapol Laongsakul

Faculty of Political Science, Rangsit University, Pathum Thani, Thailand
*Corresponding author, E-mail: chanut.n@rsu.ac.th

Abstract

This study on the injustice of artificial intelligence (AI) is qualitative research. It has two objectives: 1) to study the limitations of artificial intelligence in making judgments on human affairs, and 2) to classify the types of judgment's limitations. By employing a documentary research approach, this study utilizes the information regarding artificial intelligence with impaired judgment. The study includes AI playing roles as a job applicant screening device, an emotion recognition device, a face recognition device, and an GPT detector as a population. This research uses the concept of justice as desert, which is the cornerstone of the concept of justice, to analyze the shortcomings of AI's judgment in human affairs.

The research revealed that AI has limitations in making judgments due to its currently inadequacy to make correct ethical judgments aligned with the principle of justice as desert. AI struggles to determine individuals' uniqueness, risking racial biases. For example, AI fails to recognize black people as human beings. The injustice caused by AI is due to the biased data used in AI training. AI learns from the data sets that are predominantly composed of white males holding influential roles in society. As a result, AI might develop biases, perceiving males as superior to other genders and white individuals as superior to black people. AI employs this kind of data to make judgments in various situations. For instance, it tends to evaluate male job applicants more positively more positively than female job applicants. This is the problem of impartiality with bias. It means that although AI aims to makes unbiased judgments based on its data, that is inherently biased. As a result, AI is flawed in deciding who is worthy of what, posing a risk of being unjust.

Keywords*: *artificial intelligence, injustice, impartiality with bias

1. Introduction

Artificial intelligence (AI) refers to a machine that can think logically and carry out some tasks by autonomously. AI marks the progress of science and technology. Merriam-Webster (n.d.) defines AI as "the capability of computer systems or algorithms to imitate intelligent human behavior". AI can be classified into 3 levels according to its ability (Sittichabuncha, 2021, pp. 92-94): 1) Narrow AI is the AI that has the ability to perform specific tasks that may be equal or better than humans in those specific tasks, such as defeating the world champion in the game of Go. However, even though it has proficiency in these tasks, it lacks the versatility of humans. For example, a narrow AI that can beat specific human in the game of Go might not be able to beat the same human in other things like soccer or racing because it is only capable of specific tasks. 2) General AI, the upgraded version of narrow AI, has the ability to perform a wide range of tasks. General AI can beat humans in the game of Go and can also beat humans in soccer or racing. 3) Super AI is the

highest-level AI, exhibiting superhuman abilities in multiple areas. It is similar to general AI in that it can perform a variety of tasks but distinguished by its extraordinary proficiency in more various of task domains.

In 2024, the current year of this research, AI is still mainly operating at a narrow AI level. Several well-known AIs are designed to perform specific tasks, such as ChatGPT providing information via chat message, Google Translate doing only the task of language translation, and COMPAS assessing the risk of recidivism among offenders. However, it is undeniable that even a narrow AI is starting to play a bigger role in human society. There is even a concern about job displacement. The rise of AI includes the role in justice issues. For example, China uses AI for legal analysis issues or more specifically, employing it to act as a judge which can raise the question of ethical issues.

Ethics is the science of considering of what is right/wrong, good/bad. These are the questions of ethics that requires a great deal of discretion. The famous professor of political philosophy from Harvard University, Michael Sandel, has raised questions about whether important ethical decision-making should be delegated to AI or humans. He further argues that "AI not only replicates human biases, it confers on these biases a kind of scientific credibility. It makes it seem that these predictions and judgments have an objective status." (cited in cited in Pazzanese, 2020) This paper highlights the issue of "impartiality with bias" which could lead to negative impacts for humans.

Sharna and Graydon (2021) found that AI contains biases, including 1) sample bias, which occurs when the data used to train the AI does not present the entire population accurately. For example, ImageNet, a popular image dataset for AI training, is more likely to display more images of white people when prompted with the word "bridegroom". This is due to the fact that the dataset from ImageNet in general tends to depict images of white people over black people. Another example happened in Boston, USA where an AI named StreetBump was developed to detect and report road issues, such as potholes. However, it was discovered that areas with lower-income residents had lower road problem reports compared to areas with wealthier residents. This is because low-income people are less likely to possess smartphones and, therefore, cannot report data to StreetBump. 2) Historical bias is the bias of AI arising from actual societal biases, and it is transferred onto the AI. This type of bias can occur even when the data is the correct representative of the entire population. For example, in job applications screening, AI tends to accept male applicants more than female applicants because a company has mostly male employees, as the AI is trained on data predominantly composed of male employees. 3) Aggregation arising from the assumption that all individuals have identical qualities without considering their differences. This type of bias disadvantages minorities because the measurement is only applicable to the broader population. For example, relying solely on blood sugar levels for diabetes testing may lead to misdiagnoses among certain racial groups due to the correlation between ethnicity and blood sugar levels.

Former judge Katherine Forrest (2021) has examined in her works that AI is designed by humans and highlighting the concern that humans' biases may permeate AI. Thus, it's best not to completely rely on AI in terms of ethical judgment. Wang and Tian (2022) discussed the reliance on AI in China's legal field, which ranges from document verification to aiding in legal disputes. However, Wang also points out that AI has difficulty comparing similar legal cases and sometimes overlooks important information. Tania Sourdin (2018) asserts that while AI can assist human judges or lawyers with paperwork, humans should be the ones to make the ultimate judgment.

ChatGPT, following it launch in November 2022, had reached 100 million users by January 2023 (Hu, 2023). This rapid growth came with a risk of academic corruption as ChatGPT demonstrated proficiency in passing the academic test. It was reported that a law instructor from the University of Minnesota tested ChatGPT's capabilities by administering a law exam covering four subjects. The law instructor lacks

information to distinguish between answers belonging to AI and which answers belonging to humans. It was found that ChatGPT attained a score of C+, meeting the criteria to graduate from the Faculty of Law in some programs. However, students at the University of Minnesota have an average score of B+ (Sloan, 2023). Even though AI is capable of passing the exam, it still does not possess superior abilities to humans. This is consistent with what the focus of this paper, which assumes that AI still has limitations in its intelligence.

These issues cause the researchers to hypothesize that AI possesses limitations and flaws in using its discretion to judge humans in term of ethical affairs, leading to the research question of whether AI has limitations in using its discretion.

2. Objectives

1) To identify the limitations of artificial intelligence in terms of ethical judgment.
2) To categorize the limitations of artificial intelligence in terms of ethical judgment.

3. Materials and Methods

This qualitative research utilizes a documentary approach to study four types of AI, classified by tasks, which include:

1) AI as a face recognition device/photo analysis device
2) AI as a job applicant screening device
3) AI as an emotion recognition device
4) AI as a GPT detector

This research uses the concept of "justice as desert" to identify the limitations of AI in terms of ethical judgment. The theory states that justice is giving each person his/her due (Aristotle, 2009). Justice as desert involves value judgments consisting of three elements, developed from Wigley, 1998, as follows:

1) Deserver: This element is meant to indicate who is worthy of receiving things. This research will evaluate AI's ability to accurately identify the deserving, such as its ability as a job applicant screening device.

2) Reason: This element is meant to explain the rationale behind why a person deserves things. This research will evaluate whether AI uses the appropriate reason in decision-making processes, such as its retionale for identifying the cheater who uses ChatGPT to take exams.

3)Treatment: This element is meant to indicate what the person should receive. This research will evaluate the ability of AI's ability to determine whether the treatment the deserver receives is the correct treatment or not.

4. Results and Discussion

The research found that AI had the problem of impartiality combined with bias because it used the biased criteria to judge unbiasedly.

1) AI as a face recognition device had a problem of racism. There was a concern that AI often displayed images of white people over black people (BBC, 2020). The problem was reported by Colin Madland, a Tweeter user (right side of Picture 1) who tweeted an incident with Zoom. He reported that Zoom removed the head of his black friend (left side of Picture 1) when his friend used the virtual background feature. It could be concluded that Zoom failed to recognize black face as the face of a human. When Colin Madland posted this on Twitter, he encountered the same issue because the preview picture focused solely on the face of Colin Madland, a white man, which was not what he intended to emphasize. The preview image should have shown the face of his black friend, which was removed by the AI (see Picture 2).

Picture 1 Black people's head was removed by AI. (https://twitter.com/colinmadland/status/1307111818981146626)

Picture 2 Preview image used white people as a cover

The case of Colin Madland and his black friend exemplifies the issue of impartiality with bias when analyzed with the concept of justice as desert because AI perceived that humans were white not black, but AI lacked substantial justification for why black people should not be considered human. While the bias of white humans against black humans was evident due to hatred and oppression, AI's bias towards black humans was an impartiality with bias because AI lacks emotions. AI judged everything by existing data, which was full of white humans. When analyzed with the concept of justice as desert, the research found that black people didn't deserve the treatment of having their heads removed because AI didn't have a valid reason to do so in terms of ethical value. The finding of the case of Colin Madland and his black friend's case was similar to the research from Johns Hopkins University, Georgia Institute of Technology and University of Washington (Rosen, 2022) reveling a bias against black people and women. The test asked AI to distinguish between a houseworker, a criminal and a doctor based on the pictures. AI tended to identify women as housekeepers, black people as criminal, Latinos as janitors, and men as doctors over women. Notably, there wasn't any information in the picture to indicate any person's occupation; there were no prison uniform or a white coat. It suggests that the way AI judges people from the pictures might be influenced by sample bias and historical bias.

2) AI as a job applicant screening device had a problem of gender bias. One of the world's most famous e-commerce platforms, Amazon, used AI to screen job applicants for the company. However, it seemed that AI had the problem of gender bias because it learned from the existing data of the company's past 10 years which is predominantly comprised of male employees. Amazon's AI was able to filter out applicants that have a similar demographic profile to its own workforce. Therefore, male applicants were more likely to be selected than female applicants (Min, 2023, p. 3812). AI tended to select applicants who used typical words, used by male engineers, in their resumes, such as the words "executed" and "captured". (Dastin, 2018).

The case of AI as a job applicant screening device clearly highlighted the problem of impartiality with bias, caused by historical bias, because it was the truth that male employees were the dominant workforce at that time. When analyzed with concept of justice as desert, the research found that there was insufficient information to indicate which gender had better ability. The use of words such as "executed" or "captured" did not reflect the ability of the gender. Therefore, AI didn't have a good reason to justify that male applicants outperformed than female applicants.

3) AI as an emotion recognition device had limitations in comprehending human emotions, which could potentially lead to the issue of racism. Lauren Rhue (2018) found that when AI was tasked with evaluating emotion from images of smiling white people and smiling black people, it tended to label smiling black people as an angry person more frequently than smiling white people. This evaluation may result in peculiar and unfair recommendations for black people. For example, if black people want to be recognized as happy people, they need to smile as best as possible while white people do not have to put an extra effort. This was an unnecessary burden for black people and can be considered an injustice.

A serious injustice issue was found when this case was analyzed using AI as a job applicant screening device. A giant company, such as Unilever, used AI to evaluate the facial expressions, body language, and choice of words of the job applicants, resulting in significant time savings for the company in screening job applicant (Booth, 2019) . However, as per Lauren Rhue's finding in 2018, AI might have a flaw in evaluating human emotions. Consequently, there might be a situation where an applicant with equal or higher quality than another may not be selected due to the incorrect evaluation of their facial expressions and body language by the AI. Lisa Feldman Barrett, a professor of psychology at Northeastern University, explained that the a single human expression conveys multiple meanings, but AI was trained to interpret human expressions in a universal way (Devlin, 2020). This impartiality with bias was the result of aggregation bias, which was caused the assumption of AI that everyone had the identical qualities without considering individuals' differences. When analyzed with concept of justice as desert, the research found that AI didn't have enough reason to judge person's attitude based on facial expressions or body language.

4) AI as a GPT detector, used for detecting ChatGPT usage in students' assessments, had limitations in detecting AI-generated essays. Liang et al. (2023) uncovered that AI misjudged man-made TOEFL essays as being written by ChatGPT. Non-native speakers were misjudged by the rate of 61.3%, indicating approximately 61/100 of man-made essays were misjudged as AI-generated essays. On the other hand, if essays were written by a native speaker of English, the AI would misjudge only 5% of the essays as AI-generated essay. Essays written by non-native English speakers were more likely to be judged as written by AI than essays written by native English speakers because AI evaluation was based on whether it could predict the selection of words in essays. If the AI accurately predicted the vocabulary, it would consider it was the work generated by AI due to the uncomplicated vocabulary usage. The writings of non-native English speakers were more likely to be assessed as being generated by AI because their language skills awere more limited than those of native English speakers.

This was impartiality with bias, which was caused by aggregation bias due to the failing to the diversity of humans. Therefore, using the GPT detector would put non-native English speakers at a disadvantage, as they may be falsely accused of cheating for using ChatGPT to write their essays. When analyzed with the concept of justice as desert, the research revealed that utilizing biased measurements was unjust, irrespective of the outcome. Accurate judgment requires accurate measurement.

5. Conclusion

The research found that artificial intelligence (AI) carried the risk of being unjust in making ethical judgments due to the utilization of biased data in training it. The issue was not caused by the AI's hatred or bias towards anyone. For example, if the AI that was trained with the information indicating that a doctor was white, a criminal was black, and a maid was a woman, it would misjudge the quality and state of any individual. Thus, this paper used four types of AIs classified by their function as evidence, which were: 1) AI as a face recognition device. This AI had a problem of racism due to its training predominately on white human faces, indicating the problem sample bias and historical bias. 2) AI as a job applicant screening device had a problem of gender bias due to its training predominately on male employees, indicating the problem of historical bias. 3) AI as an emotion recognition device had the problem of racism due to its training predominately on viewing human expressions in a universal way, lacking diversity in training data. This indicated aggregation bias 4) AI as a GPT detector had the problem of favoring native English speakers because AI considered that genuine humans should be proficient in writing the complex English essays. This also indicated aggregation bias.

All the bias discussed in this paper— sample bias, historical bias, and aggregation bias, was hidden behind the guise of the neutrality of AI because AI does not have feelings and operates based on logic. The paper suggested the issue of "impartiality with bias" where AI judged everything impartially based on biased criteria from biased data. Sample bias and aggregation bias can be easily mitigated by AI developers, but the true challenge is the historical bias because it reflects the reality of the world. AI developers alone cannot change the societal reality where women are currently holding lower social status than men. Black people are currently holding lower social status than white people. This phenomenon requires everyone to consider whether it is right or wrong.

6. Acknowledgements

This research was supported by the Research Institute of Rangsit University

7. References

Aristotle. (2009). *Nicomachean Ethics* (D. Ross, Trans.). Oxford: Oxford University Press.

BBC. (2020). *Twitter investigates racial bias in image previews*. Retrieved October 5, 2023, from https://www.bbc.com/news/technology-54234822

Booth, R. (2019). *Unilever saves on recruiters by using AI to assess job interviews*. Retrieved October, 5, 2023, from https://www.theguardian.com/technology/2019/oct/25/unilever-saves-on-recruiters-by-using-ai-to-assess-job-interviews

Dastin, J. (2018). *INSIGHT-Amazon scraps secret AI recruiting tool that showed bias against women*. Retrieved October 10, 2023, from https://www.reuters.com/article/idUSL2N1WP1RO/

Devlin, H. (2020). *AI systems claiming to 'read' emotions pose discrimination risks*. Retrieved October, 5, 2023, from https://www.theguardian.com/technology/2020/feb/16/ai-systems-claiming-to-read-emotions-pose-discrimination-risks

Forest, K. B. (2021). *When machine can be judge, jury, and executioner*. Singapore: World Scientific.

Hu, K. (2023) *ChatGPT sets record for fastest-growing user base*. Retrieved January, 12, 2024, from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Liang, W., Yüksekgönül, M., Mao, Y., Wu, E. Q., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Merriam-Webster. (n.d.) Artificial intelligence. In *Merriam-Webster.com dictionary*. Retrieved October 2, 2023, from https://www.merriam-webster.com/dictionary/artificial%20intelligence

Min, A. (2023). Artificial intelligence and bias: challenge, implication, and remedies. *Journal of social research 2*(11), 3808-3817

Pazzanese, C. (2020) *Ethical concerns mount as AI takes bigger decision-making role in more industries*. Retrieved October, 5, 2023, from https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

Rhue, L. (2018). *Racial Influence on Automated Perceptions of Emotions* Retrieved October, 1, 2023. from https://ssrn.com/abstract=3281765.

Rosen, J. (2022). *Flaw AI makes robots racist, sexist*. Retrieved October, 5, 2023, from https://hub.jhu.edu/2022/06/21/flawed-artificial-intelligence-robot-racist-sexist/

Sharna, S., and Graydon, M. (2021). *Social Bias in AI and its Implications*. Retrieved April, 10, 5, from https://ntrs.nasa.gov/api/citations/20210010446/downloads/NASA-TM-20210010446.pdf

Sittichanbuncha, Y. (2021). Artificial Intelligence (AI) and Its Use in Healthcare and Emergency Medicine. *Journal of Emergency Medical Services of Thailand, 1*(1), 91-104.

Sloan, K. (2023) *ChatGPT passes law school exams despite 'mediocre' performance*. Retrieved January, 13, 2023, from https://www.reuters.com/legal/transactional/chatgpt-passes-law-school-exams-despite-mediocre-performance-2023-01-25/ [2024, January, 13].

Sourdin, T. (2018). Judge v Robot? Artificial Intelligence and Judicial Decision-Making. *University of New South Wales Law Journal*, *41*(4). https://doi.org/10.53637/zgux2213

Wang, N., & Tian, M. Y. (2022). "Intelligent Justice": human-centered considerations in China's legal AI transformation. *AI And Ethics*, *3*(2), 349–354. https://doi.org/10.1007/s43681-022-00202-3

Wigley, S. (1998). *The role of desert in distributive justice*. A thesis for the degree of Doctor of Philosophy. London School of Economics and Political Science.