International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:09, Issue: 12 "December 2024"

A Comparison of the Relative Effectiveness of AI vs Humans in Debugging Computer Code

Yuvan Ganne and John Leddo

MyEdMaster LLC
Virginia, USA

DOI: 10.46609/IJSSER.2024.v09i12.022 URL: https://doi.org/10.46609/IJSSER.2024.v09i12.022

Received: 2 December 2024 / Accepted: 23 December 2024 / Published: 28 December 2024

ABSTRACT

Current research and educational assessment have concluded that the majority of AI programs have AI bias, which causes their data to be skewed (Schwartz et al., 2022). Because of this trend that can affect almost all AI programs, it is hard to know how large of a problem AI bias can be and how much it can affect data representation. This paper explores whether AI bias exists in finding errors in computer code. We conducted an experiment in which we took 2 AI code debuggers and 2 human programmers. Next, we tested the AI code debuggers against 24 different pieces of code with various errors, using the human programmers as a control comparison. The results of the experiment showed that the AI's accuracy of error was around 95% while the humans were around 89%. These results showed that, in regard to the AI programs used for error detection in coding, AI bias is not a frequent problem that can impact data heavily.

Introduction

Over the years there have been numerous technological advances in the field of artificial intelligence. Ever since AI was established as a field in 1956 during the Dartmouth Conference, organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, it has been improved upon at an vast and ever growing rate. When first created, AI was used only for playing chess and checkers. Now, only 70 years later, AI can be used to write programs, create pictures, correct spelling, have conversations, and much more.

However, despite such technological advances, AI has had many shortcomings through the years. When in training, a skewed data set can cause some AI programs to have biased predictions. In fact, In 2023, AI algorithms led to non-white patients getting more Cesarean procedures unnecessarily. Also, In 2023, a class action lawsuit accused UnitedHealth of illegally using an AI algorithm to turn away seriously ill elderly patients from care under Medicare Advantage. The lawsuit blamed naviHealth's nH Predict AI model for inaccuracy (Horowitz, 2024). In other words, if an AI's training data and prediction modeling is skewed or incorrect it can cause grave mistakes to befall those on the receiving side of the AI's predictions.

Another example of this is that, back in 1988, the UK Commission for Racial Equality found a British medical school guilty of discrimination. The computer program it was using to determine which applicants would be invited for interviews was determined to be biased against women and those with non-European names. However, the program had been developed to match human admissions decisions, doing so with 90 to 95 percent accuracy. What's more, the school had a higher proportion of non-European students admitted than most other London medical schools. Using an algorithm did not cure biased human decision-making. But simply returning to human decision-makers would not solve the problem either (Manikiya, 2019).

That begs the question, should AI replace humans on certain tasks, or does AI bias cause AI to be too prone for errors and lacking in accurate results? In the present study, we will test whether or not AI can give the same level or more accurate results when tested against its human counterparts. In the study, we pitted 2 AI Code checkers and 2 human programmers against each other in order to see which is more accurate in detecting computer bugs, AI or humans?

Methods

Participants

The participants for this experiment consisted of 2 programmers who were fairly versatile in programming. Both participants passed two necessary criteria for this experiment. The first criterion was that they had to be over the age of 21 in order for them to participate. Another criterion was that both participants had to have at least 5 years of coding experience. Along with that, the participants had to have actively used coding at least once per month, or had a profession that specializes in coding. However, both participants had to have been around equal skill level. This allowed the data to be unbiased and not skewed toward one side or another based on the difference in skills levels between the two programmers. This experiment was volunteer only, which eliminated biased answers caused by the incentive of money.

Examples and Technology Used

The technology used for this experiment consisted of 2 AI coding error detectors.

The two sites used were the two free sites, zzzcode.ai and OpenAI's ChatGPT software. For this experiment, there were a total of 24 examples. These examples consisted of 8 simple errors, 8 complex errors, and 8 pieces of code with no errors. The errors we used for the simple examples were RTE Errors(Runtime Errors), Misspelled Var errors, Printing sentence errors , Incorrect loop errors, and LE Errors. For the complex coding errors, there were assent loading errors, complex RT errors, and pieces of code with multiple errors in them. Lastly, for the code with no error, the same examples that were used for the simple errors were reused but with the mistakes corrected.

Procedure

During this experiment, the two AI Coding Error debuggers had numerous coding language options. However, for both we chose to use Javascript as our language since it was the most common, and probably had the most data in the AI training set. For the experiment, we would take the code we chose beforehand and input it into the debugger. Afterwards, the debugger would debug the code, identifying the problem and explaining how to resolve it.

Since there are multiple ways are a variety of coding languages, ways of reading code, and debugging code we chose a simpler approach. Before inputting our code, we would make sure all the examples were in javascript since that was the most popular coding language and most likely had the most examples when the AI for the debugger was in its training phase. After collecting all the examples of code needed for the experiment, we first used the AI checkers to debug the code. When we imputed the code, if the AI checker correctly identified the error, we would mark the result for that example as 1 and if it was incorrect, we would mark it as 0. After we did this for all 24 examples, we showed the two professionals the same 24 examples, asked them to identify any errors, and scored 1 or 0 depending on whether or not they correctly identified the errors.

Results

For the codes which contained simple errors, both AI checkers correctly identified and classified all of the errors. However, the first human checker misidentified one bug and the other two missed two. For the eight pieces of code with the complex errors, the first AI checker misidentified 3 of the complex errors for the simple variant and the second AI checker misidentified 4 for its simple counterpart. Similarly, both human checkers also misidentified 3 complex errors for their simple counterpart. The next eight pieces of code had no errors among them. As an effect, both Ai checkers and both humans found no errors in all eight pieces of code. Lastly, the final eight pieces of code also consisted of eight simple errors. In those eight examples, both AI checkers did not miss an error. However, the second human checker missed an error in one of the pieces of code. These results suggest that human and AI checkers performed virtually identically.

Discussion

The primary purpose for this paper was to compare AI and Human in terms of efficiency, accuracy, and the comparative agreeability between the 2 humans and 2 AI models. After completing the testing phase of the investigation, we discovered 2 notable things. First, despite the technological breakthroughs in AI programming over the years, their results were comparable to those of their human counterparts. For the section of the experiment where they would have to identify a simple error in the code, the AI scored 100% accuracy on average. Meanwhile, the humans scored 81.25% accuracy on average. Similarly, for the part of the evaluation with complex bugs, both AI and Human checkers were 56.25% accurate on average. However, unlike the human debuggers, both AI checkers evaluated different pieces of code incorrectly. The first AI checker got the first, second, fourth, fifth, and seventh examples correct. On the other hand, the second AI checker got examples two, three, six, and seven correct.

This discrepancy in the data is most likely caused by over-inflation in the training data which caused skewed results. To clarify, if an AI program receives a multitude of one type of problem and not much of any other, it could cause the AI to skew the results in a shift towards the type of problems it has experience with. Lastly, both AI checkers never missed a bug entirely or detected a non-existent bug. Similarly, the humans also never found a bug in clean code. However, they did miss a bug entirely making their accuracy score slightly worse than the AI in that section. They scored a 93.75% while the AI had 100% accuracy. In total, the humans were 89.06% accurate while the AI checkers were 95.94% accurate.

These results show that AI, when being compared to humans, show little to no bias in their training data set when analyzing and correcting mistakes. As a result, for this type of task, AI can be used without significant bias. This suggests that bias in AI may be task dependent and depend on the quality of the training data. More research is needed to articulate when bias is likely to occur in AI and how to mitigate it.

References

Agbolade Omowole. (2021, July 19). Research shows AI is often biased. Here's how to make algorithms work for all of us. World Economic Forum. https://www.weforum.org/stories/2021/07/ai-machine-learning-bias-discrimination/

Henderson, S. (2022, March 16). There's More to AI Bias than Biased Data, NIST Report Highlights. NIST; NIST. https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

How AI Bias Is Impacting Healthcare. (2024). Informationweek.com.

https://www.informationweek.com/machine-learning-ai/how-ai-bias-is-impacting-healthcare#

IBM Data and AI Team. (2023, October 16). AI Bias Examples | IBM. Ibm.com; IBM. https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples

Moschella, D. (2022, April 25). AI Bias Is Correctable. Human Bias? Not So Much. Itif.org. https://itif.org/publications/2022/04/25/ai-bias-correctable-human-bias-not-so-much/

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns, 2(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, 1270(1270). https://doi.org/10.6028/nist.sp.1270

Tyson, A., Pasquini, G., Spencer, A., & Funk, C. (2023, February 22). 60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care. Pew Research Center Scienociety. https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care

What do we do about the biases in AI? | McKinsey. (n.d.). Www.mckinsey.com.

https://www.mckinsey.com/mgi/overview/in-the-news/what-do-we-do-about-the-biases-in-ai\