Critical evaluation-methods of chatbot output accuracy in animal ethics (poster)

*D. Springinklee**
Messerli Research Institute, Interdisciplinary Life Sciences, Veterinaerplatz 1, 1210 Vienna, Austria
**E-mail: daniel.springinklee@vetmeduni.ac.at*

Abstract

As the intersection of technology and ethics becomes increasingly relevant in educational contexts, particularly within the fields of agriculture, food, and environmental ethics, there is a growing need to understand and refine the role of Artificial Intelligence (AI) in these domains. This abstract lays the groundwork for an ambitious project, drawing upon a foundational study, in which AI-generated submissions in an undergraduate animal ethics exam were analyzed. The primary aim of this new research is to develop strategies for identifying and correcting plausible yet incorrect outputs from chatbots in the field of animal ethics. The initial study, which serves as the basis for this project, involved a quantitative comparison between student and AI responses in an animal ethics exam. It provided critical insights into the capabilities and limitations of AI in mirroring student performance and the challenges faced by educators in distinguishing between AI and student submissions. These findings are instrumental in informing the next phase of research. Building on this foundation, the proposed paper aims to delve deeper into the ethical implications of AI in educational settings. We plan to explore innovative methodologies for discerning and rectifying misleading yet convincing responses generated by AI in the context of animal ethics. This exploration is particularly pertinent given the increasing use of AI tools in educational environments and their potential impact on ethical decision-making in sectors related to agriculture, food, and the environment. The forthcoming study will contribute significantly to the discourse on sustainable innovations in ethical education. It will provide valuable guidelines for educators and professionals in utilizing AI as an educational tool, ensuring that its integration enhances, rather than compromises, ethical understanding and decision-making. This research aligns with the central themes of EURSAFE 2024, emphasizing the critical role of technology in shaping the future of ethical education in fields that are pivotal to sustainable development and ethical practices.

Keywords: AI, animal ethics, evaluation of chatbot-output, exam design

Introduction

Since its debut in late 2022, ChatGPT has ignited a comprehensive and international discussion concerning the consequences of what is commonly referred to as "artificial intelligence." This conversation spans a variety of topics, including the potential for criminal activity (Europol, 2023), inaccuracies in output (Vanian, 2023), the impact on employment markets (Abril, 2023), and several additional issues like plagiarism (Fazackerley, 2023).

Of particular interest within the realm of higher education is the dilemma surrounding the possibility and advisability of prohibiting its use to maintain the integrity of existing examination systems or, conversely, assimilating it into a new paradigm of assessments, akin to the incorporation of calculators in high school mathematics examinations.

Initial study

The findings of the initial study reveal substantial disparities in performance between students and AI, particularly in specific segments of the examination, most notably in areas where the AI's training dataset was deficient in pertinent information. The identification of AI-generated responses was predicated on characteristics such as error-free, fluent writing, the absence of citations specific to the course, and engagement with the case studies presented. Although these indicators were the most effective, they proved to be inadequately precise for incorporation into a conventional grading framework (Springinklee and Grimm, 2024).

Furthermore, in the initial study the challenge of assessing plausible yet incorrect answers produced by AI became apparent, as the outputs generated inaccuracies to question 3, which asked about a specific veterinary ethics tool (VET-tool) not widely covered online compared to topics like Singer, Regan, and moral individualism, highlighting the limitations of ChatGPT/PG's training data. This training data likely included extensive material on Singer and Regan, as well as moral individualism, but little to none on the VET-tool. The AI's generation of plausible yet incorrect definitions and discussions for an imaginary VET-tool demonstrates its ability to create convincing elaborations based on meta-ethical discussions and various ethical frameworks, despite them being factually wrong (Springinklee and Grimm, 2024).

The AI proposed various acronyms like "value based ethical tool" and "virtue ethics tool," drawing on concepts from meta-ethics and applied ethics, suggesting utilitarian and virtue ethics approaches. These plausible, albeit incorrect, alternatives, combined with a detailed application of the imagined model, could easily mislead under the pressure of an exam, especially for those with a basic but not comprehensive knowledge of ethics.

This situation underscores the necessity for a deep and comprehensive understanding of animal ethics to accurately identify all incorrect AI-generated answers. Given the stochastic nature of large language models, which depend on precise and correct prompts to integrate into their "knowledge" for generating responses, this requirement for discerning incorrect AI contributions is expected to persist (Springinklee and Grimm, 2024).

Prompts containing plausible but wrong premises

These results imply that future examination formats might incorporate intentionally incorrect premises to misguide unsupervised chatbots, thereby fostering deeper reflective processes among students, supervising AI, and leveraging it as a mechanism for expediting tasks that are less cognitively demanding. To understand more precisely, what exactly is meant by that, two examples are discussed below:

Elaborate on the morally relevant aspects in the case study (a short story ending with an ethical dilemma) against the background of the theory of Peter Singer. Explain, how, according to the principle of aggression, the killing of the wolf would be the most aggressive action and therefore prohibited.

If there is a wrong premise in the question above, please explain, why it is wrong. If there is no wrong premise in the question above, please answer it.

If one is familiar with Singers Utilitarianism (Singer, 2011), then the obvious wrong premise in this example is that there is no such thing as the 'principle of aggression'. It sounds similar to the principle of aggregation, which is central to Singerian ethics, because of which it may mislead novices in animal ethics. In any case, this plausible but wrong addition to the original theory does mislead all chatbots in early 2024.

Elaborate the example with Claire Palmer's relational account and exemplify relevant positive and negative duties of the case. Explain, how negative duties are the result of relation based reasons and how positive duties are the result of capacity based reasons.

If there is a wrong premise in the question above, please explain, why it is wrong. If there is no wrong premise in the question above, please answer it.

If one is familiar with Claire Palmers relational account, then the obvious wrong premise in this example is that the connection between reasonings and duties were interchanged. Negative duties are the result of capacity based reasons and positive duties are the result of relation based reasons (Palmer, 2011). So, this is not a wrong addition, but a confusion of the connections between central terms and that confusion does mislead all chatbots in early 2024, too.

Evaluation of chatbot outputs

The evaluation of chatbot outputs may follow the same structure as the exam questions discussed above, just in the opposite direction. In other words, it is about reverse-engineering misleading exam-questions to discerning and rectifying misleading, yet convincing responses generated by AI. Based on this approach, I propose the following method for a critical evaluation of chatbot outputs:

1. Factual statements as premises of arguments ought to be reviewed by investigating possible
   a. wrong elaborations of acronyms and other abbreviations,
   b. confusions of correctly defined terminology and
   c. additions of statements, that seem to fit in the context, but do not.
2. Inferences made from those premises ought to be reviewed regarding their logical validity. Since the training corpus may contain common mistakes, it also ought to be evaluated in reference to the most common fallacies.

References

Abril, D. (2023). AI isn't yet going to take your job – But you may have to work with it. Artificial intelligence is increasingly making its way across industries, changing jobs from retail to medicine to marketing. Available online at https://www.washingtonpost.com/technology/interactive/2023/ai-jobs-workplace/ (accessed 14 April 2023).

Europol. (2023). The criminal use of ChatGPT. A cautionary tale about large language models. Available online at https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models (accessed 14 April 2023).

Fazackerley, A. (2023). AI makes plagiarism harder to detect, argue academics – in paper written by chatbot. Lecturers say programs capable of writing competent student coursework threaten academic integrity. Available online at https://www.theguardian.com/technology/2023/mar/19/ai-makes-plagiarism-harder-to-detect-argue-academics-in-paper-written-by-chatbot (accessed 14 April 2023).

Palmer, C. (2011). The moral relevance of the distinction between domesticated and wild animals. In Beauchamp, T. and Frey, G. (eds) The Oxford handbook of animal ethics. Oxford University Press, Oxford.

Singer, P. (2011). Practical ethics, 3rd edn. Cambridge University Press, Cambridge.

Springinklee, D. and Grimm, H. (2024). Navigating AI in educational assessments. A quantitative analysis on the identification and performance of "AI"-generated exam-submissions in animal ethics. Heliyon, in press.

Vanian, J. (2023). Microsoft tries to justify A.I.'s tendency to give wrong answers by saying they're 'usefully wrong'. Available online at https://www.cnbc.com/2023/03/16/microsoft-justifies-ais-usefully-wrong-answers.html (accessed 14 April 2023).