Case Studies in AI Ethics

Coursework 2

Ethical Considerations in AI Content Moderation: A Case Study of OpenAI's Partnership with Sama

B243694 - 08

1 Utilitarianism (max 1 page)

1.1 Brief description of theory

Utilitarianism focuses on the results, or consequences, of our actions, and treats intentions as irrelevant. Good consequences indicate good actions, and according to Bentham[1] and Mill[2], and even the Greek philosopher Epicurus, actions should be measured based on the happiness or pleasure they bring[3]. AI ethics that use utilitarianism entail optimising AI systems for the highest good of society. This method, emphasizing good consequences (like happiness, pleasure, welfare, benefits ...) for the majority, guides the design of algorithms while taking stakeholder impact, transparency, and fairness into account. It adheres to moral guidelines for the responsible development of AI and seeks to guarantee beneficial contributions to society[4].

1.2 Application to case study

From the perspective of utilitarian ethical analysis, the contractual arrangement between OpenAI and Sama for content moderation services is a complex situation that requires a thorough assessment of its effects on the welfare of society. Based on the maximisation of overall happiness or utility, utilitarianism offers a framework for evaluating the moral implications of this arrangement by weighing the benefits and drawbacks for each stakeholder.

Through the partnership between OpenAI and Sama, harmful content will be filtered out of the digital ecosystem, protecting consumers from harm. Since this goal strengthens user trust and platform integrity while also making the online community safer, it is consistent with the utilitarian principle of promoting the greatest good. Additionally, the programme creates jobs in areas where there are few other work possibilities, which may help reduce poverty[5]. These favourable results point to a rise in societal utility that benefits many different stakeholders, such as platform users and the local areas where jobs are created.

But this arrangement also raises important moral questions, especially in regards to the well-being of Sama's Kenyan employees. According to reports, these people experience psychological distress as a result of the type of content they moderate, which is made worse by poor pay and unfavourable working conditions[5]. Justice and equity are called into question by the payment structure's glaring economic inequality, which shows that OpenAI paid Sama a substantial amount more than the workers' salaries. Moreover, the project's ostensible social benefits are undermined by the workers' instability and well-being due to the lack of job security, which is made worse by the abrupt termination of contracts[6][7].

The ethical assessment of this situation from a utilitarian perspective necessitates a careful examination of how these results impact total utility. The negative effects on the workers' mental health and financial situation indicate a serious ethical shortfall, even though the project's objectives and some of its results may favourably benefit society welfare. Utilitarianism's principles demand that proactive steps be taken to prevent these unfavourable effects in addition to their anticipation. This entails making certain that workers receive just compensation, offering sufficient mental health care, and promoting open, moral, and ethical workplace environments.

To maximise society benefit, the OpenAI-Sama content moderation initiative is currently being carried out in a way that fails its utilitarian objectives. Although this project has the potential to improve digital safety and boost economic growth, it fails ethically because of the serious injuries that the workers suffer. Substantial reforms are necessary to prioritise the welfare of these people in accordance with utilitarian principles. These reforms should include improvements to working conditions, mental health care, and compensation. Such modifications are essential to ensure that the project truly serves the greatest good by benefiting all interested parties, as well as to ensure an equitable allocation of advantages and liabilities. Only then will the project be able to realise its potential for beneficial societal effect.

2 Deontology (max 1 page)

2.1 Brief Description of theory

Deontology, especially Kantian Ethics, places more emphasis on obligation and the inherent morality of deeds than on their effects. It is predicated on Immanuel Kant's[8] formulation of the categorical imperative. According to this ethical paradigm, deeds are ethically correct if they can be universalized, meaning that anyone can carry them out, and if they view humanity as a goal in itself rather than just a means to an end. In order to decide under this framework, more details would be required, such as whether the action respects the autonomy of each and every person impacted and whether it could be consistently willed as a universal law.

2.2 Application to case study

Deontological ethics, in particular Kantian ideas, provide a unique viewpoint on the ethical considerations in the case of OpenAI's collaboration with Sama for content moderation. According to Kant's categorical imperative, people must be treated with respect for their intrinsic autonomy and dignity, not just as means to an end but as ends in and of themselves. This ethical framework places a strong emphasis on the value of moral obligations and rights, no matter what the outcome.

Examined from a Kantian perspective, the agreement between OpenAI and Sama presents serious ethical questions. First off, it may be argued that the work environment and nature of the job—which exposes workers to potentially unpleasant content without providing enough mental health support—treat Kenyan workers like tools. This instrumentalization of people goes against the Kantian command to respect each person's inherent worth and dignity, mainly for the sake of enhancing AI performance and protecting digital spaces.

The unequal compensation, where employees are paid far less than the industry average, along with the unstable employment that results from contracts ending suddenly, compounds the moral conundrum[5][6][7]. Such actions violate the fundamental principles of Kantian autonomy by restricting the workers' capacity to make wise decisions and follow their own objectives.

A disregard for the wellbeing and mental health of the workers is also evident in the insufficient notice given to them of the nature of the content they were to moderate and the reportedly ineffectiveness of the ”wellness” counselling sessions they were given[5]. This mistake exposes a failure to treat these people with the respect and regard they need, as rational agents capable of suffering and deserving of help, in addition to neglecting the duty of care that OpenAI and Sama owe to their employees.

According to Kant, a more moral strategy would call for deeds that uphold the autonomy and dignity of the employees. This entails offering efficient mental health support, creating open and moral working environments, and guaranteeing fair compensation that is in line with the demands of the job and its influence on mental health. Furthermore, actions like providing frequent breaks, regularly auditing working conditions, and aggressively promoting counselling service use would be consistent with the Kantian need to consider people as ends in themselves, thereby fostering their autonomy and well-being.

In conclusion, when analysed through the prism of deontological ethics, the case study of OpenAI's outsourcing of content moderation to Sama exposes serious ethical flaws. To ensure that the workers are treated as autonomous individuals with inherent dignity rather than as simply tools for obtaining technological improvements, a dramatic reorganisation of the project would be necessary in order to adhere to Kantian ideas. The project can only hope to meet the strict requirements of deontological ethics and respect the moral rights and duties inherent in the employer-employee relationship by taking such an ethical approach.

3 Meta-Analysis (max 1 page)

Through a synthesis of the arguments put forth under utilitarian and deontological frameworks, my position attempts to traverse the complicated ethical concerns of OpenAI's decision to outsource content moderation to Sama. This case study exposes a convergence of ethical issues that go beyond the scope of a single ethical theory and necessitate a comprehensive approach.

The benefits of content moderation in improving user experience and online safety are highlighted by the utilitarian perspective, which emphasises the greatest good for the greatest number of people. It recognises that workers in areas with a dearth of formal employment have access to economic opportunities that could spur regional economic growth. However, this point of view needs to address the serious injuries done to the workers, such as psychological suffering and exploitation, which call into question the utilitarian estimate of the total benefit to society.

Deontological ethics highlights the moral duty to regard Kenyan workers as ends in and of themselves rather than as merely means to an end because of its emphasis on the intrinsic dignity and rights of individuals. This viewpoint highlights the ethical requirement of treating people fairly and providing appropriate help for those exposed to potentially traumatic content, criticising the project for failing to respect the autonomy and inherent value of the workers.

Using these frameworks as a guide, my position tends towards a hybrid ethical position that values worker welfare and dignity while acknowledging the advantages of content regulation for society as a whole. The hybrid should also involve other more compassionate approaches to content moderation such as the capability approach and care ethics. The capability approach, which emphasises personal empowerment for well-being and dignity, and care ethics, which emphasise empathy and the value of relationships, can work together to build supportive environments that acknowledge the duties of moderators and the psychological difficulties they face [9][10]. This position is important because it advocates for a balanced strategy that makes sure that technical gains do not come at the expense of human well-being, while also acknowledging the complexity of ethical considerations in technology and AI research.

The literature on the development of ethical AI supports this position by highlighting the significance of taking into account both the rights of those involved in the production of the technology and its effects on society [11]. AI ethical principles, like the ones put forth by the IEEE Standards Association, place a strong emphasis on openness, responsibility, and giving human welfare top priority when designing and implementing AI systems [12]. Other businesses, like Facebook, have employed a mix of in-house and external moderating techniques to mitigate the psychological impact that content moderators endure proving that there are ways to lessen the harms [13][14].

It is possible to find counterarguments that the potential for technical improvement and the economic rewards outweigh the labour sacrifices. This viewpoint, however, ignores the moral duty to protect the wellbeing and dignity of every person concerned, especially those who are in vulnerable situations.

If policies were put in place to greatly enhance the content moderators' pay, benefits, and mental health support (such as those create by facebook), I might change my mind. In order to reduce ethical issues and bring the project closer to both utilitarian and deontological standards, it may also be necessary to ensure meaningful consent and transparency regarding the nature of the activity.

The ethical study of OpenAI's outsourcing to Sama concludes by highlighting the necessity of a balanced strategy that upholds human rights and dignity while pursuing the wider societal advantages of AI technology. This case study is an essential reminder of the moral difficulties that arise when AI is used and the need to handle these issues with caution and morality.

4 References

[1] J. Bentham, *The Principles of Morals and Legislation.* Amherst, NY: Prometheus, 1789.

[2] J. S. Mill, *Utilitarianism*, G. Sher, Ed. Indianapolis, IN: Hackett, 1861.

[3] J. Driver, "The History of Utilitarianism," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Winter 2022, Metaphysics Research Lab, Stanford University, 2022.

[4] S. Liao, *Ethics of Artificial Intelligence.* Oxford University Press, Incorporated, 2020, ISBN: 9780190905033. [Online]. Available: https://books.google.co.uk/books?id=2ST3DwAAQBAJ.

[5] B. Perrigo and J. Zorthian, "How OpenAI's ChatGPT labeling of data in Kenya sparked ethical outrage," *Time*, Jan. 2023. [Online]. Available: https://time.com/6247678/openai-chatgpt-kenya-workers/.

[6] Africapay, *Minimum wages in Nairobi, Kenya*, 2024. [Online]. Available: https://africapay.org/kenya/salary/minimum-wages/2182-cities-nairobi-mombasa-and-kisumu.

[7] K. N. B. of Statistics, *Minimum wages in Kenya*, Trading Economics, latest year. [Online]. Available: https://tradingeconomics.com/kenya/minimum-wages#:~:text=Minimum%20Wages%20in%20Kenya%20remained,Kenya%20National%20Bureau%20of%20Statistics.

[8] I. Kant, *Groundwork of the Metaphysics of Morals*, 1998th ed., trans. by M. Gregor. Cambridge: Cambridge University Press, 1785, Translated and edited by Mary Gregor, with an introduction by Christine M. Korsgaard.

[9] C. Koggel and J. Orme, "Care ethics: New theories and applications," *Ethics and Social Welfare*, vol. 4, no. 2, pp. 109–114, 2010.

[10] I. Robeyns and M. F. Byskov, "The Capability Approach," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Summer 2023, Metaphysics Research Lab, Stanford University, 2023.

[11] B. Mittelstadt, "Principles alone cannot guarantee ethical ai," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, Nov. 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0114-4. [Online]. Available: https://doi.org/10.1038/s42256-019-0114-4.

[12] "Ieee recommended practice for assessing the impact of autonomous and intelligent systems on human well-being," *IEEE Std 7010-2020*, pp. 1–96, 2020. DOI: 10.1109/IEEESTD.2020.9084219.

[13] Innodata, *The ethics of content moderation: Who protects the protectors?* https://innodata.com/the-ethics-of-content-moderation/, Accessed: 2023-03-17, 2022.

[14] C. Papaevangelou and N. Smyrnaios, "The case of a facebook content moderation debacle in greece," in *Journalism and digital content in emerging media markets.* Cham: Springer International Publishing, 2022, pp. 9–26.