Outline & Abstract on Online Content Moderation

Salvin Chowdhury

March 27, 2025

The Outline

1 Requirements of the Outline & Abstract

The purpose of this paper is to craft the outline and abstract with regards to the right to content moderation. We look over the outline requirements and the structure of the paper, and then develop an argument for and against the thesis statement.

1.1 Outline Requirements

The requirements of the paper are to have a title, an abstract that culminates in a thesis statement, a rough outline, and a list of three properly formatted sources as listed below:

- Reading from the class
- Entry from the Stanford Encyclopedia of Philosophy
- A source of choosing

1.2 What a Good Philosophy Paper Should Look Like

A good philosophy paper should present a clear thesis and argue for it. The argument needs to be persuasive and presented in a well-ordered, logical fashion. We then consider the strongest possible objections to the view and offer rebuttals. Basic fallacies should be avoided, and appropriate quotations should be used to clarify points.

1.3 Structure of Paper

The paper should have the following structure:

- An introduction and conclusion
- A body that presents the arguments/evidence for the thesis
- One or two strong arguments against the thesis

1.4 Sources of Research

Avoid using Wikipedia, YouTube, and any random website. Examples of good sources are peer-reviewed sources, the Stanford Encyclopedia of Philosophy, Google Scholar, JSTOR, Philosopher's Index, and books.

1.5 Abstract Structure

The first paragraph should be focused heavily on. Avoid using general introduction statements. Be sure to set a brief introduction by letting the reader know the topic and why the topic is relevant. The reader should be given a play-by-play account of how the essay is going to unfold and include a thesis statement.

1.6 Assumptions of the Reader

Assume that the reader is a jerk and overly hostile. The job is to convince the reader while respecting their intelligence. Always give the opposition the best reading and don't assume the reader will do the same. Be sure to explain the concepts fully, and ensure that anyone should be able to read the paper and understand it.

1.7 Crafting the Thesis Statement

The thesis statement should come as the last sentence of the first paragraph. The thesis statement should be true or false.

2 Introduction, Title & Abstract

In this paper, I discuss about the what the paper is going to be about, the title of the paper, as well as how the abstract of the paper is going to look like.

2.1 The Introduction

This paper discusses in detail about content moderation on the internet. We look at the parties that have been historically involved in moderating content, and whether the moderation of the content should be done by humans that represent an entity that is in charge of the online platform or whether policing of speech should be done by done by training an artificial intelligence model that does all the work.

2.2 Title of the Paper

As per the introduction, I wanted the title to reflect what the introduction discusses about. Hence, the title of the paper would be:

"Who should decide the rules for content moderation and should it be automated using Artificial Intelligence?"

2.3 The Abstract

As the debate to who has the right to remove and moderate content rages on, the different societal views on what may seem morally wrong and right greatly differ at the same time. This makes it a great challenge for those in charge, whether it is private companies, governmental institutions or even artificial intelligence, on whether they should have the right to hold the reins of controlling the free speech that goes on to grow at a exponential rate on the internet. This paper explores cases similar to LambdaMOO and decisions made by different groups on free speech. We argue that it is private companies, who should hold the reins of controlling the free speech on the internet due to a lack of political bias, however we argue that it is better to have humans moderating content rather than automating it using Artificial Intelligence (AI).

3 Structure of the Body Paragraphs

In this section, I discuss about the body of the paper and the kind of content that will be discussed in each body of the paragraph. The goal is to simply discuss about arguments that support and go against the thesis statement. I would like to also expand upon the arguments in detail the demonstrate the strength of these arguments.

3.1 Paragraph One: The Early Internet & Moral Philosophies

In this paragraph, we look at case studies of incidents that occurred within communities on the online internet. We discuss what decisions were made as a result of such incidents, and who made these decisions and why. We also create a connection using these case studies with the moral philosophies from Immanuel Kant & Thomas Hobbes.

The purpose of this paragraph is to demonstrate the real life impacts on human beings when they engage in online conversations on the internet via a online account.

3.1.1 A Rape in Cyberpsace & Immaneul Kant

A Rape in Cyberspace is an article by Julian Dibbell. The article discusses about a incident that had occurred on LambdaMOO, which is a multi-user dimension program that allows users to create their own characters and perform interactions with other characters using just text. One night, a user, "Bungle" had decided to take control of two other characters and had performed sadistic actions on them. Although the harm was done virtually, the action had translated into physical harm in the form of mental distress for the humans who were in charge of the virtual characters.

Later on, knowledge of the incident was known by the entire community, sparking outrage and a demand for justice. In a group meeting with all users on the platform, some had advocated for "Bungle" be removed while others stated that the user had done nothing wrong as the platform had no established rules at the time. While the arguments raged on, the moderators had decided at the end to just remove "Bungle" from the platform. This article presents how members of a online community argue about what is justice to them, as well as negative effects of virtual harassment. Some of the key takeaways from the article was as follows:

- Virtual behavior can have real psychological and physical consequences. The controllers of virtual characters have responsibilities for those consequences
- The developers and controllers of a virtual environment interface have responsibilities for the wellbeing of the humans that control an online account on a online platform

Here, we can create a connection using these key takeaways with the philosophies of Immanuel Kant with regards to the dicussion of perfect and imperfect duty, as well as the duties of online users and moderators to have good intention when it comes to interacting with other online users.

3.1.2 Searching for a Leviathan in Usenet & Thomas Hobbes

Usenet is a computer conferencing network where users can send private messages to one another via email. Usenet has no central authority which monitors access or control, instead, it is done at the site level in the form of communities. I use this paper to create a connection about online users existing and how they interact on a online platform with other users with the philosophy of Thomas Hobbes. These are the following connections to be discussed:

- How Usenet users on the internet have a "personae", which is a personality that a user may embody on a online platform. For example, if a user makes jokes constantly, they may have a "humorous" personality
- How Usenet users can exert power on other user's using text-based dialogue. For example, using aggressive language or attempts at winning a online argument.

Here, we use the philosophy of Thomas Hobbes to demonstrate that online interactions are more than just conversations, that they have impacts on the real world. For example, if a online user is persuasive at promoting left wing ideology and can influence other online users, this will impact the real world in the sense that the humans behind the influenced online accounts will express support for such ideologies.

3.2 Paragraph Two: Who Gets to Moderate?

After we have established that online interactions on the internet can real consequences, we now look at who should have the right to moderate over interactions and content on the internet. In this paragraph, we look at arguments from different papers on who should moderate the internet and why.

The purpose of this paragraph is to demonstrate and present the strength of the arguments as to why and why not a certain entity should not be allowed to moderate online content.

3.2.1 Government & Content Moderation

The CEO of Cloudflare, Matthew Prince, suggested that the government should be more involved in deciding what speech is allowed online. But the article argues that it is a bad idea because private companies have been in charge of regulating speech on their platforms. If the government had more power over online content, it could threaten free speech and independence of tech companies.

Tech companies are working on handling harmful content. The paper argues that even though their efforts aren't perfect, they are better suited for the job. If the government started controlling what content is allowed, it could become a political tool to suppress speech that official don't like. This would be dangerous as it could lead to censorship based on political agendas rather than fairness.

3.2.2 Moderation by the Platform or by the Online Community?

This paper discusses about the the two different approaches to moderating content on the internet. The first is the moderating of content by the platform itself, this means the technical administrators who run the platform and have the power to remove users at will. The second is the moderation of content by online moderators themselves who don't represent the platform but represents the community they moderate for.

The paper details the two perspectives in depth, which is the platforms and policies perspective and the communities perspectives. The platforms and policies persepctive focuses on content moderation from the perspectives of online social platforms, and the community perspective focuses on volunteer community moderators who are actually more akin to community leaders.

3.3 Paragraph Three: The Argument on Content Moderation by AI

After we have established the perspectives of different entities being able to moderate content on the internet, we now look at if that moderation should be automated by Artificial Intelligence and the papers that support and go against AI moderation.

The purpose of this paragraph is to demonstrate and present the strength of the arguments as tho why and why not the entity in charge should useartificial intelligence should and should not be able to moderate the internet.

3.3.1 Content Moderation using AI

Automating content may often be justified due to the sheer size of content on the internet. It is desirable to use AI due to the immense amount of data , the relentlessness of violations and need to make judgments without human moderators making them.

However, the problem is that the AI tools simply compare new posts to previously flagged content. This means they're great at catching duplicates but not great at identifying new forms of harmful speech. Such AI tools also struggle with understanding context, sarcasm and cultural meanings. AI systems also make mistakes that can unfairly impact marginalized groups.

As such systems work based on patterns in large amounts of data, they unintentionally reinforce existing biases. For example, they may wrongly flag certain communities more often while failing to protect them from real harm.

3.3.2 Fighting Hate Speech, Silencing Drag Queens?

This article discusses about how AI tools used for moderating content don't fully understand the context, and this may negatively impact the LGBTQ community. For example, some LGBTQ people, especially drag queens, use words that AI systems might label as toxic, however they may be actually playful or empowering. However, because such words are used as insults in general conversations, AI tools may assume them as being always offensive.

4 The Arguments

In this section, I discuss about the arguments that will be presented in the final version of the paper, as well as the counter arguments. These ideas are discussed in the detail to present the strength of these ideas.

4.1 Modarting of Content by Government

Here, I explore the arguments as to why governments should be able to moderate content on the internet. I also explore the counterarguments as to why governments should not have the right to moderate. As a note, I assume that a government consists of officials who have been elected by their respective constitutuents in a democratic process. An example of a government would be the United States government.

4.1.1 Governments Should Not Moderate Content

Governments should not be able to moderate online content as people who are elected into government represent political bias. People who have political biases may have positions on issues, and those positions may seem moral to them, a position that may not be shared by groups of inidividuals who oppose such a position. As a result, such a bias may be used to silence dissidents who otherwise express dissent against the elected official's views on critical issues.

4.1.2 Governments Shouldn't Moderate Content

Governments should be able to moderate content as they consist of officials who have been elected in a democratic manner. As the responsibility of the government is to uphold the free speech laws that is enshrined in the constitution, the people who have the power to uphold rests in the hands of those elected officials. It would not make sense for a unelected representative to be able to make decisions on content moderation. We justify this using the philosophy of Thomas Hobbes, supporting the argument using ideas such as "Social Contract" where if inidividuals should adhere to the constituion in real life, then they should also do this online. Otherwise, it would lead to a "State of War".

4.2 Moderating of Content by Platorms

IHere, I explore the arguments as to why platforms should be able to moderate content on the internet. I also explore the counterarguments as to why platforms should not have the right to moderate.

4.2.1 Platforms Should Moderate Content

Platforms, such as Facebook, should be able to moderate content because if the government had power over online content, it would threaten free speech and the independence of technology companies. Technology companies are better suited to handle harmful content than government employees, and have the necessary resources to make it happen. If the government were able to control what content is allowed, it would become a political tool to suppress free speech. This would be dangerous as it could lead to censorship.

4.2.2 Platforms Shouldn't Moderate Content

Platforms shouldn't be able to moderate content because they may be owned by a board of directors or a chief executive officer (CEO) who may have political biases. Furthermore, if there are no existing laws that prevent the influence of such platform executives through donations or bribes, such executives can be easily influenced to control speech and alter it to make it more favorable to the donor. For example, if a political candidate bribed a social media executive to suppress speech that is critical of the candidate and promote speech that favors the candidate, it may result the candidate winning the election as online users are influenced to believe that the political candidate they saw on the internet would be a better choice for office than their opponent.

4.3 Moderating of Content by Volunteer Moderators

Here, I explore the arguments as to why volunteer moderators should be able to moderate content on the internet. I also explore the counterarguments as to why such volunteer moderators should not have the right to moderate. As a note, a volunteer moderator is someone who volunteers to moderator and is not a respresentative of the platform they moderate on. A volunteer moderator is rather a moderator of a online community.

4.3.1 Volunteer Moderators Should Moderate Content

Volunteer Moderators should be able to moderate content as the prime motivation behind moderating a community is their passion for the community they're a part of. Such moderators can't be influenced by any external parties, hence their motivation to moderate remains true. Furthermore, the survival of a online community depends on the quality of the volunteer moderator, as bad moderating could lead to a decrease in interactions within the online community. In other cases, bad moderating can lead to removal of a volunteer moderator and be replaced by a new one by the will of the users of the online community.

4.3.2 Volunteer Moderators Shouldn't Moderate Content

Volunteer Moderators shouldn't be able to moderate content because they are vested with the power to censor any user that may express opinions against them. Their power to censor comes with no consequences, as the user being censored can no longer access the community they've been removed from to express the injustice that has been comitted against them. As a result, good moderation will depend on the mercifulness and maturity of the moderator, something which can't be guaranteed to be found in all moderators.

References

Dias, O. T., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online [Copyright - © Springer Science+Business Media, LLC, part of Springer Nature 2020; Last updated - 2024-03-26]. *Sexuality & Culture*, *25*(2), 700–732. https://msoe.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/fighting-hate-speech-silencing-drag-queens/docview/2495185828/se-2

Gillespie, T. (2020). Content moderation, ai, and the question of scale [Copyright - © The Author(s) 2020. This work is licensed under the Creative Commons Attribution – Non-Commercial License https://creativecommons.org/licenses/by-nc/4.0/ (the "License"). Last updated - 2024-11-17]. *Big Data & Society*, *7*(2). https://msoe.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/content-moderation-ai-question-scale/docview/2473716357/se-2

Hobbes, T. (2018). Xiii: Of the naturall condition of mankind, as concerning their felicity, and misery. In *Leviathan*. Lerner Publishing Group.

Huff, C., Johnson, D., & Miller, K. (2002). Virtual harms and virtual responsibility: A rape in cyberspace. *IEEE 2002 International Symposium on Technology and Society (ISTAS'02). Social Implications of Information and Communication Technology. Proceedings (Cat. No.02CH37293)*, 323–330. https://doi.org/10.1109/ISTAS.2002.1013833

MacKinnon, R. C. (1992). *Searching for the leviathan in usenet* [Doctoral dissertation, ProQuest Dissertations and Theses] [Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-02-20]. https://msoe.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/searching-leviathan-usenet/docview/304027876/se-2

Prichard, H. A. (2002). Kant's fundamental principles of the metaphysic of morals. In *Moral writings*. Oxford University Press.

Samples, J. (1972). Why the government should not regulate content moderation of social media, 2019, april 9. *Retrieved from CATO Institute: https://www. cato. org/publications/policy-analysis/why-government-should-not-regulate-contentmoderation-social-media (2.12. 2020)*.

Seering, J. (2020). Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW2). https://doi.org/10.1145/3415178