2021

Can deep learning be used to make Discord communities safer?

IT2B, INHOLLAND UNIVERSITY OF APPLIED SCIENCES

FRISART, ZOEY

JUNE 6, 2021

Acknowledgements

First, I would like to thank my lecturers Thijs Otter, Inge Wisselink and Murian dos Reis Ribeiro for providing me with the necessary knowledge and tools to be able to perform this research.

Furthermore, I would like to thank Petra Delsing, Quartermaster Artificial Intelligence at “Ministerie van Infrastructuur en Waterstaat” for pear reviewing and providing feedback on this report. The input and feedback gained here were very useful and interesting and improved the quality of my work.

I would like to thank Guido van Dijk and Manoah Tervoort for pear reviewing this paper, and providing me with in depth feedback on the paper. This allowed me to improve the overall quality of the report.

I would like to also thank Richard Frisart and Esther Smeenk for proofreading this report and allowing me to throw my ideas and thoughts at them for a meaningful discussion.

Finally, I would like to thank Kyan and Remi for providing me the emotional support throughout the process of writing this report. It has greatly helped me with staying motivated and pushing this report to a higher level.

Summary

This report analyses the use of deep learning for making Discord communities safer, from a legal and social-ethical perspective. The research is performed through ‘desk-research’, with the research type being ‘exploratory research’. It uses the legal and ethical cycle as the two main tools for analysing. The report finds that explicit consent from the user should be gained, and that transparency is key in the successful use of deep learning. It finds that it is ethical to use deep learning to ensure the safety of the community. It advises that it is implemented according to the ALTAI self assessment.

Table of contents

Introduction

Context

Discord communities enable people with a shared interest to connect. To foster a healthy Discord community care should be given to ensuring the safety of the users. As such the content in the Discord community should be moderated. Traditionally this is done by a team of moderators in the Discord community. While this approach might work for small Discord communities, this approach does not scale up so easily. For large Discord servers with several hundred or even thousands of members, moderating all content becomes increasingly difficult for moderators.

This is where deep learning could assist moderators by flagging potentially harmful messages to moderators. Deep learning reduces the amount of content moderators need to review. And allows them to focus their efforts on resolving incidents instead of monitoring all channels constantly.

Research questions

Main question: Can deep learning be used to make discord communities safer?

1. Ethical sub-question: Is it morally acceptable to use machine learning in a Discord community, to ensure the safety of the members?
2. Legal sub-question: What are the legal implications of processing all messages in a Discord community?
3. Legal sub-question: In the current context of Discord, bots do not need consent of the users in a community. Is it legal to process the data in the circumstance that you do not have an explicit consent?

Motivation / relevance

I am part of a moderator team in a growing Discord community. This specific Discord community includes a lot of members that are at risk of toxicity and/or harassment, as it contains a lot of members of the LGBTQI+ community.

For this reason the moderation team takes extra care in ensuring the safety of all members, this however becomes increasingly difficult for a moderation team of 3 members. To tackle this problem the start of development of a Discord bot that moderates the content of Discord community using deep learning has been made.

During the development and (current) pilot phase, questions/concerns were raised by some members in the community. As the moderation team takes these concerns seriously, this has motivated this research.

Research goal / objectives

This paper aims to identify possible problems that the development and deployment of a Discord bot like this could pose. From a legal point of view, it should raise legal issues that should be considered. From an ethical point of view, it should identify if the use of machine learning to process and profile messages and/or users is ethically justifiable to be able to ensure the safety of the general public.

The findings will be used by the moderation team of the “Panda Squad” Discord community to determine if and how the team could use deep learning to ensure the safety of the community.

Reading guide

Introduction: Introduces the context of the research, the motivation, the goals and how to read this paper.

Approach: Explains the research methods used during this research.

Legal analysis: Analyses our legal sub questions, using the Legal cycle. It analyses several articles of the GDPR, and an analysis of a verdict imposed by CNIL. This section focusses on legal complications related to the input required for the deep learning.

Social-ethical analysis: Analyses our ethical sub question using the Ethical Cycle. This section aims to determine if the use of deep learning is ethically acceptable.

Conclusion: States the conclusion and recommendations made during the legal and social-ethical analysis.

Bibliography: Credits all the sources used for this research paper.

Appendix: Reflection on the competence 'Research skills'

Glossary

To ensure that the information is clear, the following terminology is used in the report.

- **ALTAI** Assessment list for trustworthy artificial intelligence
- **Discord** An application that provides a place to talk and voice chat with other people, similar to Microsoft Teams and Slack.
- **Discord bot** A robot user that extends the functionality of a discord community.
- **Discord channel** A text or voice channel that is about a specific topic within a Discord community.
- **Discord community** A group/server within Discord, these can be joined by an invite link.
- **Discord moderator** A person that is responsible for ensuring the rules of a Discord community are enforced. They are responsible for fostering a safe and welcoming community.

Approach

The research approach of this paper is desk research. The type of research is exploratory research. It aims to identify possible problems/hypothesis rather than proofing a hypothesis.

Sources will be found using search engines, google scholar and databases of universities and high schools for earlier studies for this purpose. To determine if a source is reliable I will look at who wrote it, validate that the information is still relevant and why the information was published and who the target audience of the information was.

All sources used will be stated in the bibliography. This research will use documents published by Discord, as Discord is a key part of this research. As Discord is however not a neutral party, information they provide is reviewed from a critical point of view.

For legal analysis this paper aims to look into the terms of service of Discord, GDPR and conclusions of lawsuits that might be applicable. The legal aspect of this paper is based upon the European Union. The following search terms are used to find sources for the legal part “GDPR”, “What falls under personal data”, “What falls under legitimate interest”, “Discord developer policy”, “GDPR article 7”, “GDPR enforcement”, “GDPR conditions for consent”, “GDPR article 9”, “GDPR automated decision making” and “Conditions for manifestly made public”.

For the legal analysis the ‘Legal Cycle’ by Thijs Otter will be used. For the ethical analysis the ‘Ethical Cycle’ by I. van de Poel & L. Royakkers will be used.

Legal analysis

This chapter will dive into the legal considerations that go into the processing of data within a discord community. As this would be a basic condition for using machine learning on it.

What are the legal implications of processing all messages in a Discord community?

To use deep learning to make a Discord community safer would require a Discord bot that analysis all messages that are send within a Discord community. To answer our research questions we need to first determine what possible implications could arise in regards to processing all messages.

As such our legal question is “What are the legal implications of processing all messages in a Discord community”.

Which legal area applies

Which rules and legislation apply will be key to answering our research question, this however greatly differs per country. For this question we will mainly look at privacy concerns, as the Discord community is operated from within the European Union, we will focus on European law related to privacy. While additional laws may apply based on the purpose of which the data is processed, we will mostly limit the research for this question to the privacy laws.[1]

Which rules apply?

On May 25, 2018 the European Union has put into effect the General Data Protection Regulation which from now on will be referred to as the GDPR (GDPR, 2018). The GDPR puts into place strict rules in regards to processing personal data.

[1] As the research is related to Deep learning, it briefly mentions the concept regulation proposed by the European Commission This paper however does not go into depth about this legislation.

For the processing of the messages that are send in the community the administrators of a Discord community can add a Discord bot to the community. This gives the Discord bot access to the Discord community through the Discord API. For this the Discord bot is obligated to comply with the Discord terms of service, privacy policy, developer policy and developer terms of service. Although these are not laws they are a legal agreement between Discord and the bot developers.

Discord policies

Let's first analyse the requirements and restrictions that Discord puts into place, as these can serve as a base for analysing the GDPR in relation to Discord bots.

The Discord developer terms of service state in Section 2a that "You will comply with all applicable privacy laws and regulations including those applying to personally identifiable information ("PII")." (Discord, 2020, "User Privacy And Security" section). This clause in the developer terms of service requires the Discord bot to comply with the GDPR, as this is the applicable privacy in the region from where the Discord bot will operate.

In the Discord developer policy is stated that "You may not retain data any longer than necessary for the operation of your application" (Discord, 2020). This is in line with Article 5 of the GDPR paragraph 1e which states "Personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed" (GDPR, 2018, Chapter 2, Article 5, para. 1e).

While the Discord bot does not directly store messages for a prolonged period of time, messages that are flagged are send to moderators. As discord stores messages indefinitely, this results in the personal data related to flagged messages to be stored indefinitely.

The Discord developer policy further states that "You may not process Discord data in a way that surprises or violates Discord users' expectations" (Discord, 2020). Discord goes into slightly more detail to meaning in their support article "Bot Verification and Data Whitelisting".

They state you should ask yourself the question “Would someone be surprised by this?” (Discord, 2021).

GDPR

As the Discord bot will use machine learning to flag messages to moderators, we should also consider Article 22 of the GDPR which grands additional rights to data subjects, which in this case would be members of the Discord community, in regards to “Automated individual decision-making, including profiling” (GDPR, 2018, Chapter 3, Article 22). Section one defines “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling” (GDPR, 2018, Chapter 3, Article 22, para. 1).

In Article 22 of the GDPR section two it states that “Paragraph 1 shall not apply if the decision is based on the data subject’s explicit consent.” (GDPR, 2018, Chapter 3, Article 22, para. 2c). This will be further covered in the legal sub-question “In the current context of Discord, bots do not need consent of the users in a community. Is it legal to process the data in the circumstance that you do not have an explicit consent?”

While we aim to keep the scope of this research question to privacy law, it is worth mentioning that the European Commision has published a concept regulation in regards to Artificial Inteligence. This regulation would impose transparency obligations to the Discord bot as it would fall under Title IV as the Discord bot would interact with humans (European Commision, 2021, “TRANSPARENCY OBLIGATIONS FOR CERTAIN AI SYSTEMS (TITLE IV)” section).

In Article 9 of the GDPR restrictions are set in place in regards to “Processing of special categories of personal data” (GDPR, 2018, Chapter 2, Article 9). While processing all messages in a Discord community does not directly aim to process special categories of personal data, it might unintentionally do so if members of the Discord community have a conversation about these subjects.

As paragraph 1 states “Processing of personal data revealing racial or … pollical opinions … shall be prohibited.” (GDPR, 2018, Chapter 2, Article 9, para. 1). In paragraph 2e it provides an exception to paragraph 1 if “Processing relates to personal data which are manifestly made public by the data subject” (GDPR, 2018, Chapter 2, Article 9, para. 2e). This raises the question, whether sending a message in a Discord community is considered as making information manifestly public.

In article 9 paragraph 2a it gives another exception to paragraph 1. It states that “the data subject has given explicit consent to the processing of those personal data for one or more specified purposes” (GDPR, 2018, Chapter 2, Article 9, para. 2a). As such getting explicit consent from the user for processing the data could allow us to process the data as long as it is for a specific purpose that the community member has given consent for.

Advice

Based on the information presented, the advice is that the following actions should be taken. Furthermore, some possible issues have become clear and should be taken into consideration and/or further investigated.

- A clear privacy policy should be created that informs the user what data the Discord bot collects and what will it be used for.
- Research should be performed for a valid method of getting consent for the processing of data. This will be further discussed in the chapter “In the current context of Discord, bots do not need consent of the users in a community. Is it legal to process the data in the circumstance that you do not have an explicit consent?”
- Further research should be done to determine whether when the Discord bot unintentionally parses data, if that would fall under Article 9 of the GDPR. If so, it would require the Discord bot to comply with the requirements of Article 9.

The moderation team would benefit from performing a careful analysis of the data they process and for what purpose. The processing of all messages and the associated user data

that is send along with each message raises serious questions about the validity of processing these without explicit consent.

There are most definitely some legal issues in processing all messages in the current context.

In the current context of Discord, bots do not need consent of the users in a community. Is it legal to process the data in the circumstance that you do not have an explicit consent?

Now that we have established some possible problems in regards to processing of all messages, a lot of these problems where related to consent and informing the end user about the data collection. Discord currently does not require consent of all the users in a community for a bot to process the data, this research sections aims to answer “Is it legal to process the data in the circumstance that you do not have an explicit consent?”.

Which legal area applies?

As states in the section “What are the legal implications of processing all messages in a Discord community?”, we will mainly look at privacy laws from the European Union.

Which rules apply?

To determine the rules that apply we will look at the Discord policies and the GDPR. As these are the two most relevant to the research question.

On August 20, 2017 a new Discord developer terms of service went into effect, which has one notable change as it introduced section 2.4, where Discord states “If you have access to End User Data through the API or the SDK, you shall ensure that your Applications do not collect, use and/or disclose End User Data except… If you have access to End User Data through the SDK, you additionally agree to get express permission from the End Users.” (Discord, 2017, Chapter 2, para. 2.4). This clause would require bots to get explicit consent from end-users, even though Discord provided no way to do this through Discord itself.

In the latest version of the Discord developer terms of service they have removed this clause. Instead they now require that you provide a privacy policy to your users, which clearly describes what you do with the data. They however set no requirements for how you inform the end users about this privacy policy (Discord, 2020, Chapter 2, para. a.).

The GDPR which went into effect May 25, 2018, requires explicit user consent for processing personal data. In Recital 32 Conditions for consent it states “Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication” (GDPR, n.d., “Recital 32” section).

Is the data collected personal data?

To deteremine wether the data the Discord bot receives are Personal data we would have to look at what GDPR defines as personal data. In article 4 of the GDPR it states: “any information relating to an identified or identifiable natural person” (GDPR, 2018, ‘personal data’ section). It further states “an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number…” (GDPR, 2018, ‘personal data’ section).

As the Discord bot receives data including but not limited to username, user id and profile picture. (Discord.js, 2021), it would qualify as personal identifiable data. Which means it is subject to the GDPR regulations.

Does the Discord policies provide ground for processing by bots?

As the Discord bots process personal data, it should have consent to process the data. To determine if the bots need to gain consent we need to see if the policies of Discord cover consent for bots so that the bots do not need to gain individual consent.

When looking at the conditions of consent, we can determine that consent should be informed and unambiguous (GDPR, 2018, Chapter 2, Article 7). The Discord privacy policy would not be able to cover consent for all bots as there is no way Discord can inform the users of all the data bots collect and what they use it for. As such the user consenting to the terms of service and privacy policy of Discord would not cover the consent for the processing of data by Discord bots as the user would not be able to give informed consent as they at the point of giving consent are unable to know what future bots will use their data for.

Does being a member of the community count as consent?

Recticle 32 of the GDPR further covers another important condition: “Silence, pre-ticked boxes or inactivity should not therefore constitute consent.” (GDPR, n.d., “Recital 32” section). This condition requires that consent should be explicitly given. As such being an active member of the Discord community wouldn’t be explicit consent, because the user hasn’t given us explicit consent for the processing of data by our bot.

Could we cover the processing by the Discord bot under ‘legitimate interest’?

Then we should also consider whether the Discord bot could process the user data based on legitimate interest.

According to the GDPR, “one can process personal data based on legitimate interest, in case this is necessary to perform normal, predictable activities related to the purpose of the service.” ((European Commission, 2018) (GDPR, 2018)). In other words, the user should not be surprised that data is processed. However, the legitimate interest is not useful in case the interest is overridden by the interest or fundemental rights and freedoms of the user.

The European Commisin further states that “Your company/organisation must inform individuals about the processing when collecting their personal data” (European Commission, 2018). As such the moderation team should make sure that a privacy policy is available to the Discord community members.

While the use of the Discord bot might fall under ‘legitimate interest’, the processing of the specific data that is necessary for the well functioning of the bot (e.g. all content of chats), should be considered as “overriden by the interests or fundemental rights and freedoms of the user” (GDPR, 2018, Chapter 2, Article 6) and would still require explicit consent. As not processing this type of data is almost impossible within a Discord community, it is unlikely that the Discord bot would be covered by ‘legitimate interest’.

How have these rules been applied?

On January 21, 2019 the CNIL found that Google was in violation of the GDPR, as the consent was not validly obtained. CNIL stated two reasons for this, the first being “the restricted committee observes that the users’ consent is not sufficiently informed.” (CNIL, 2019). They further observed that “the collected consent is neither “specific” nor “unambiguous”.” (CNIL, 2019).

The CNIL found that the end-user was not sufficiently informed, because the information was “diluted in several documents and does not enable the user to be aware of their extend” (CNIL, 2019). It further found that “the purposes of processing are described in a too generic and vague manner, and so are the categories of data processed for these various purposes” (CNIL, 2019). As such the user would not be able to make an informed consent and so Google was in violation of Article 6 of the GDPR.

Advice

Based on the information presented above, we advice the following actions to be taken.

A clear privacy policy should be implemented that informs the Discord community members, including but not limited to what data is collected, for what purpose, when this data is collected and who this information is shared with.

Research should be done in how explicit consent could be gained for Discord bots, as just being a member of the server and continuing to use the server would not meet the conditions of consent set by the GDPR.

As the current implementation that is actively being used in the Discord community does not have a privacy policy, there is no way the user would have been able to give an informed consent. Furthermore even if we would consider the data collected on grounds of legitimate interest, the moderation team should still inform the community members about the processing of the data. As the current implementation does not do this, the data collected is collected without a legal basis to do so.

As such this report recommends that the collected data is deleted, and that the Discord bot gets turned off until it matches the legal requirements for processing the data. The research did not whether there is a legal obligation to inform the members of the community about this, but from an ethical perspective, this seems correct.

Social-ethical analysis

This part of the research will consider whether or not the use of deep learning is morally acceptable. Because while something might be legal, that does not necessarily make it ethical to do it.

Case: Is it morally acceptable to use machine learning in a Discord community, to ensure the safety of the members?

We first need to determine what our moral problem is. The moderation team wants to ensure a safe and welcoming community for all members. As a community grows, manual moderation won't be a feasible solution alone. Machine learning could assist the moderation team by flagging potentially harmful messages to the moderation team.

Our moral problem here is that the Discord community members might consider this as an invasion of their privacy. Furthermore they might not want to be subject to profiling and automated decision making by AI. On the other hand, both the members and the moderation team want to have a safe and welcoming community space.

Furthermore, the AI might profile users based on their history within the server, and as such might flag messages from people who have been flagged before earlier than other community members. This could be for example interpreted as profiling on character traits.

The stakeholders or participating parties of this problem are the moderators of the Discord community, as they wish to use the bot for ensuring a safe community for everyone. On the other hand the members of the Discord community wish to have privacy, and because the Discord community has members that are vulnerable, they might have a negative association with profiling.

The European commission has released a report on ethics guidelines for a trust worthy AI. In this they mention seven key requirements. With regards to this, there are two points that need to be carefully considered.

- **Transparency:** “The data, system and AI business models should be transparent… Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned” (High-Level Expert Group on AI, 2019)
- **Diversity, non-discrimination and fairness**: “Unfair bias must be avoided, as it could have multiple negative implications” (High-Level Expert Group on AI, 2019)

In 2019 the “Ministerie van Economische Zaken en Klimaat” published a report “Strategisch actieplan voor Artificiële Intelligentie”. This report is only available in Dutch. It states that “the trust in AI is critical for an AI to be useful”[2], they further state that “people see the chances AI brings, but that some applications seem questionable”[3]. They further state “To maintain trust in AI it is crucial that it is used people-oriented and contributes to the welfare and well-being of people.” (Ministerie van Economische Zaken en Klimaat, 2019, Page 43).[4]

From the perspective of the Discord community members, they want to be able to trust the AI and its judgement. However for this to be possible, transparancy and communication is key. However from the perspective of the moderators, giving insight into the working of the AI gives malicious persons the ability to find ways to work arround the AI.

As the AI will use the messages of the community to learn, how do we make sure that it does not develop unfair bias? As the moderation team tries to ensure a safe community, how do we find a balance between fostering a healthy community and providing a place for free speech and meaningful discussions.

---

[2] Paraphrased and translated from “Strategisch actieplan voor Artificiële Intelligentie”, original text “Vertrouwen van burgers en bedrijven in AI is noodzakelijk voor een succesvolle ontwikkeling en toepassing van AI.” (Ministerie van Economische Zaken en Klimaat, 2019)
[3] Paraphrased and translated from “Strategisch actieplan voor Artificiële Intelligentie”, original tekst “Uit onderzoek blijkt dat burgers en ondernemers allerlei kansen zien voor AI, zoals het versterken van veiligheid, of het verbeteren van de zorg, maar ook veel twijfels hebben over de toelaatbaar-heid van sommige AI-toepassinge” (Ministerie van Economische Zaken en Klimaat, 2019)
[4] Translated from “Strategisch actieplan voor Artificiële Intelligentie”, original text “Om het vertrouwen in AI te behouden is het cruciaal dat AI mensgericht is en zo wordt ingezet dat het bijdraagt aan welvaart en welzijn.” (Ministerie van Economische Zaken en Klimaat, 2019)

Options for actions

Now that we have analysed the moral problem, we can suggest a few possible solutions. In a black-and-white-strategy this would come down to either do not use machine learning or do use it. The problem at hand however isn't that simple. And as such those solutions would do a disservice to all parties involved.

A possible solution could be being very transparent about the actions the AI has taken to the Discord community members, Discord does this by releasing a transparency report every 6 months where they go through what type of reports they received and how and why they handled it a certain way (Discord & Nelly, 2021).

Another possible solution would be to allow members to opt-out of contributing to the learning model of the AI. This way their messages are only checked but not used to further advance the AI. As such it would be less of an invasion of privacy. This however could negatively impact the effectiveness of the AI.

To ensure that the AI does not develop an unfair bias, a possible solution could be to put checks in place that tests the judgements the AI made. And if unfair bias was found, to be transparent about this and clearly communicate this to the members of the community. Going even further and reaching out to people who have been affected by this.

For the AI to be effective, it is critical that the members of the community trust the AI and its judgement. To ensure that the AI is trustworthy, the moderation team should implement the principles defined by ALTAI[5] (Ala-Pietilä, et al., 2020). If the members see that the AI is effective, and trust worthy they are more likely to be okay with the partial invasion of privacy.

[5] The assessment list for trustworthy artificial intelligence for self assessment

Ethical judgement

This section aims to look at the moral problem from the different ethical theories,

Utilitarianism

Utilitarianism determines if an action is right or wrong, based on the outcome it produces. Outcomes that result in happiness are considered good, while outcomes that result in bad things such as pain are considered bad. (Nathanson, n.d.).

L. Royakkers states that “The utilitarian framework selects the option that brings the greatest good for the greatest number” (Poel & Royakkers, 2007). With this understanding, we can analyse the moral issue and determine the most suitable approach.

The AI could limit the amount of harmful content that is available within the community. This content could cause pain and/or unhappiness for members of the community. And as such this could be considered bad according to Utilitarianism. As such the AI leading to the fast removal of this type of content would prevent the hurt. Furthermore that bad-actors are held accountable for their actions could result in happiness for community members, as their identity and wellbeing is being protected.

We should however not overlook that being flagged by the AI and/or having your voice silenced could cause pain for members, especially if the flag was unfair.

Considering these points, from an Utilitarianism point of view, the usage of machine learning to ensure the safety of the Discord community members would we morally justifiable. Because it reduces the amount of pain caused to the most people, as such the happiness it causes for most would outweigh the pain it causes to some.

Virtue ethics

The virtue approach “argues that ethical actions should be consistent with ideal human virtues” (Bonde, et al., 2013, “The Virtue Approach” section), as such when making a decission it should be done from the ideal human virtues.

The moderation team aims to create a safe environment that is welcoming and open to anyone. Their goals and motivation are in line with these virtues, because they believe that the content that the AI would flag, would not be done from an ideal human virtues. As an ideal human virtue wouldn't make you behave inappropriately against other people.

With this framework, an approach where the AI acts according to ideal human virtues would most likely be chosen. Furthermore, an approach where clear and honest communication is done about the actions and reasoning of the AI would most likely be favourable, as this would be an ideal human virtue.

Deontological ethics

Deontological ethics are based around what is right, based on a set of moral rules. "deontological theories might draw attention to the moral importance of promises, rights and obligations" (Poel & Royakkers, 2007).

Looking from this perspective, the action performed by a bad-actor would be wrong. As insulting people generally speaking isn't morally right. As such an AI that would limit harmful content would be considered good.

On the other hand, the AI is basically spying on the members of the Discord community, which would be morally wrong. The AI is furthermore build upon the principle that people won't follow the rules of the server, which assumes that people act with ill-intent.

From this point of view, the most appropriate solution would be to allow people to opt-out, as this gives them the right to not be spied on. With the assumption that people will follow the rules.

The common good approach

This framework argues that "The best society should by guided by the "General will" of the people" (Bonde, et al., 2013, "The Common Good Approach" section), it further empathisis "respect and compassion for others, especially those are more vulnerable." (Bonde, et al., 2013).

With this understanding, when making a decision the moderation team should act based on what the majority of the members want. As such, if the majority of the members are against the use of AI then the moderation team shouldn't implement AI in the community.

This framework however, also empathises protecting the vunerable. As such an AI that has a clear focus on protecting the vunerable, and enabling a community that respects each other and is welcome to everyone is the right thing.

From this perspective, the majority of the members should support the AI. For this the members should be able to see the benefit that the AI introduces. This would require the members to trust the AI. This would require the moderation team to be transparant both about the decisions it makes, and also why it makes the decisions.

As such the most appropriate action would be to implement a clear and transparant communication about the purpose of the AI, with a report once every few months that showcases how many messages the AI flagged, for what the messages where flagged and on how many of those flags the moderation team acted.

My own judgement

In this section, I will state what my action would be and why. Considering the information at hand, I would personally implement the AI to ensure the safety of the community.

Because I have closely interacted with the community myself, I have seen the damage harmful content has done to the members. Especially when it is targeted at a specific person. As such I believe that the greater good of the many outweigh the harm for some. I believe that allowing harmful content to stay in the server, fosters more harmful content as the moderation team indirectly complies with the harmful content being okay.

While I deeply care about privacy of myself and others, I do believe that ensuring the safety outweighs the invasion of privacy. I do however believe that great care should be taken in how the AI is implemented and how the communication about the AI is done. If this is failed to be done then there is a risk that it harms the people it aims to protect.

I would choose to implement the AI according to the ALTAI standard, as this would ensure the AI to be trustworthy.

Reflection

Now that we have compared the different ethical frameworks, we need to reflect on each framework and the arguments they provide.

Utilitarianism comes short in not considering justice, as while protecting the majority of the server might lead to happiness, the profiling and silencing of members could be considered as unjust. Furthermore, from a legal standpoint a person should have the right to object against automated decision making (GDPR, 2018, Chapter 3, Art. 22). Utilitarianism would argue that the hapiness the AI provides to many members would make the action morally desirable.

While Utilitarianism provides a good consideration, I wouldn't think it is sufficient to determine the moral value, as it ignores whether or not it is legal or justifiable to perform the action.

Virtue ethics comes short as it does not consider the results of the action, it purely considers if the decision is made from ideal human virtues. Because this framework fails to consider the consequences of an action it would come short in moderating a Discord community. As with moderation you need to also take into account the impact an action might have, not only on the people involved but also the surrounding community. As such I do not believe this framework would be sufficient in determining the moral value of an action.

Deontological ethics is based on the principle of moral rules, this is where it comes short as moral rules differ between cultures. As such it fails to consider the difference between cultures. As a Discord community is made up out of members all over the world from different cultures, making a judgement based on moral rules falls short as not only could that be different for each member. Even within the moderation team there are people from

different cultures. As such moral rules might fall short as they can differ within the moderation team.

The common good approach falls short in the sense that it can be based on misinformation. In case people have a bias against AI, or the moderators are not providing enough information to the end users , this could lead to them not trusting the AI. It furthermore expects that people put effort in looking into the action, which isn't always the case. A final shortcoming is that it fails to consider that people might not be able to make a judgement of what the consequence of an action would be. However as it emphasises respect and empathy to others it does closely align with the goals of the moderation team and the Discord community members.

Conclusion

In conclusion, the processing of all messages in a Discord community brings some legal complications. These complications are mainly related to consent. And while parts of the processing might fall under ‘legitimate interest’ the moderation team would gain benefits out of gaining explicit consent.

We further concludes that the data the Discord bot has collected so far does not have a legal basis to be collected on. Even if it would fall under ‘legitimate interest’, this would require the Discord bot to inform the community members about the data collection and the purpose of it. As this is currently not done, we advice that the bot gets turned off until it complies with the GDPR, and that data collected before will be erased properly.

The report concludes that the usage of deep learning would be morally acceptable, as the advantages for the majority outweigh the possible harm for the few. It however recommends that the moderation team is transparent about decisions, the actions of the AI and the data that is collected. The moderation team would benefit from implementing it according to the ALTAI self assessment. A continuous conversation with the members of the community is key to ensure the continuous understanding and fulfilling of the needs of the community.

In short, transparency is of paramount importance for the successful implementation of deep learning for the purpose of protecting the community. If members can’t trust the AI, it poses the risk to harm the people it is meant to protect.

Bibliography

Ala-Pietilä, P., Huhtamaki, S., Bauer, W., Fraunhofer, Goodey, J., Heintz, F., . . . Stix, C. (2020). *The assessment list for trustworthy artificial intelligence for self assesment.* High-Level Export Group on AI. Publications Offfice of the European Union. doi:10.2759/002360

Bonde, S., Firenze, P., Green, J., Grinberg, M., Korijn, J., Levoy, E., . . . Weisberg, L. (2013, May). *A framework for making ethical decisions*. Retrieved from Brown University: https://www.brown.edu/academics/science-and-technology-studies/framework-making-ethical-decisions

CNIL. (2019, January 21). *The CNIL's restricted committee imposes a financial penalty of 50 Million euros against GOOGLE LLC | CNIL*. Retrieved from CNIL: https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc

Discord. (2017, August 20). *Discord developer terms of service*. Retrieved from Github: https://github.com/discord/discord-api-docs/blob/b9edace323c9df64c79f104d85984690ae4e2977/docs/Legal.md

Discord. (2020, July 1). *Discord developer policy*. Retrieved from Discord Developer Portal: https://discord.com/developers/docs/policy

Discord. (2020, July 15). *Discord developer terms of service*. Retrieved from Discord: https://discord.com/developers/docs/legal

Discord. (2020, May 7). *Terms of Service*. Retrieved from Discord: https://discord.com/terms

Discord, & Nelly. (2021, April 3). *Discord Transparency Report: July — Dec 2020 - Discord Blog*. Retrieved from Medium: https://blog.discord.com/discord-transparency-report-july-dec-2020-34087f9f45fb

Discord, L. a. (2021, April). *Bot Verification and Data Whitelisting*. Retrieved from Discord support: https://support.discord.com/hc/en-us/articles/360040720412-Bot-Verification-and-Data-Whitelisting

Discord.js. (2021, April 02). *Message*. Retrieved from Discord.js documentation: https://discord.js.org/#/docs/main/stable/class/Message

Discord.js. (2021, April 02). *User*. Retrieved from Discord.js documentation: https://discord.js.org/#/docs/main/stable/class/User

European Commision. (2021, April 21). *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. Retrieved from European Commision: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Commission. (2018, August 1). *What does 'grounds of legitimate interest' mean?* Retrieved from European Commission: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/grounds-processing/what-does-grounds-legitimate-interest-mean_en

GDPR. (2018, November 14). *Art. 22 GDPR – Automated individual decision-making, including profiling*. Retrieved from GDPR.eu: https://gdpr.eu/article-22-automated-individual-decision-making/

GDPR. (2018, May 25). *Art. 4 GDPR – Definitions*. Retrieved from GDPR.eu: https://gdpr.eu/article-4-definitions/

GDPR. (2018, November 14). *Art. 5 GDPR – Principles relating to processing of personal data*. Retrieved from GDPR.eu: https://gdpr.eu/article-5-how-to-process-personal-data/

GDPR. (2018, May 25). *Art. 6 GDPR - Lawfulness of processing*. Retrieved from GDPR.eu: https://gdpr.eu/article-6-how-to-process-personal-data-legally/

GDPR. (2018, May 25). *Art. 7 GDPR - Conditions for consent*. Retrieved from GDPR.eu: https://gdpr.eu/article-7-how-to-get-consent-to-collect-personal-data/

GDPR. (2018, November 30). *Art. 9 GDPR – Processing of special categories of personal data*. Retrieved from GDPR.eu: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited

GDPR. (n.d.). *Recital 32 - Conditions for consent*. Retrieved May 25, 2021, from GDPR.eu: https://gdpr.eu/Recital-32-Conditions-for-consent/

High-Level Expert Group on AI. (2019, 04 08). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. Retrieved from European Commission: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Ministerie van Economische Zaken en Klimaat. (2019). *Strategisch actieplan voor Artificiële Intelligentie.* https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie/Rapport+SAPAI.pdf: October.

Nathanson, S. (n.d.). *Utilitarianism, Act and Rule*. Retrieved May 27, 2021, from Internet Encyclopedia of Philosophy: https://iep.utm.edu/util-a-r/

Otter, T. (2017). *The Legal Cycle.*

Poel, I. v., & Royakkers, L. (2007). *The Ethical Cycle.* Journal of Business Ethics. doi:10.1007/s10551-006-9121-6

Appendix

STARR	EXAMPLE
SITUATION	Writing a research paper on the usage of deep learning to make Discord communities safer, for legal and social-ethics.
TASKS	I was tasked with performing a legal and social-ethical research into the topic.
ACTION	I approached this task by performing the research related to the research question, making use of the information that was taught during the lectures.
RESULT	A report that I'm proud of and is of high quality, that will be used to determine how deep learning will be implemented in the actual Discord community.
REFLECTION	I have a passion for the topic I choose, as such I was motivated to perform the task at hand. In the future I would limit myself to 1 legal sub-question as while I believe the information gained in the second legal question was very useful, it also lead to me going over the word limit by quite a bit.