Ai-contentanalyse

Ai-gestuurde moderatie functie om misbruik tegen te gaan, zonder dat het in strijd gaat met het doel van de app Telegram.

Created by: Wanida
Created on: March 21, 2025 1:28 PM
Changed on: March 22, 2025 1:59 PM

Impact on society

What impact is expected from your technology?

This category is only partial filled.

What is exactly the problem? Is it really a problem? Are you sure?

Het probleem met Telegram is dat vrijheid van meningsuiting, vaak wordt misbruikt. Het gebrek op controle leidt vaak tot haatdragende berichten en criminaliteit in groepschats. Dit is een probleem, omdat het modereren in grootschalige groepen lastig is. De technologie "ai-contentanalyse" wil haatzaaien, criminaliteit en misinformatie tegengaan, wat schadelijk is voor de gebruikers.

Are you sure that this technology is solving the RIGHT problem?

This question has not been answered yet.

How is this technology going to solve the problem?

This question has not been answered yet.

What negative effects do you expect from this technology?

This question has not been answered yet.

In what way is this technology contributing to a world you want to live in?

This question has not been answered yet.

Now that you have thought hard about the impact of this technology on society (by filling out the questions above), what improvements would you like to make to the technology? List them below.

AI-contentanalyse kan verbeterd worden door de gebruikers van Telegram meer controle te geven over de moderatie-instellingen. Verder moet het systeem transparant werken en er een optie zijn om bezwaar in te dienen, als inhoud onterecht wordt verwijderd door AI. Hierdoor behoud je de gebruikers die onterecht gecensureerd worden.
Het detectiesysteem moet de privacy beschermen van de berichten, door middel van versleutelingen te gebruiken in de chat.

Hateful and criminal actors

What can bad actors do with your technology?

This category is only partial filled.

In which way can the technology be used to break the law or avoid the consequences of breaking the law?

Door de technologie van AI-contentanalyse in te schakelen, kunnen gebruikers de Ai-systeem proberen te verleiden door haatzaaiende berichten te camoufleren, zodat het niet door de algoritme wordt gedetecteerd. De uitdaging met deze technologie is een goede balans te vinden, omdat te veel controle kan leiden tot het beperken van meningsuiting. Anderzijds, te weinig laat ruimte open voor het misbruiken van de chat.

Can fakers, thieves or scammers abuse the technology?

This question has not been answered yet.

Can the technology be used against certain (ethnic) groups or (social) classes?

This question has not been answered yet.

In which way can bad actors use this technology to pit certain groups against each other? These groups can be, but are not constrained to, ethnic, social, political or religious groups.

This question has not been answered yet.

How could bad actors use this technology to subvert or attack the truth?

This question has not been answered yet.

Now that you have thought hard about how bad actors can impact this technology, what improvements would you like to make? List them below.

Het systeem kan verbeterd worden door de gebruiker de mogelijkheid te geven om zelf te kunnen instellen wat voor inhoud wel of niet gefilterd moet worden door AI. Door bijvoorbeeld bepaalde prompts in te voeren, schadelijke afbeeldingen of haatdragende zinnen, hierdoor worden detectie accurater en meer op de gebruikers zijn voorkeur afgestemd.

Privacy

Are you considering the privacy & personal data of the users of your technology?

This category is only partial filled.

Does the technology register personal data? If yes, what personal data?

Ai-contentanalyse kan persoonlijke gegevens registreren, zoals berichten en gebruikersinteractie om de inhoud van de chat te monitoren. Het creëert een algoritme door haatdragende berichten of misinformatie te analyseren op schadelijke inhoud. Dit helpt bij het sneller detecteren van dit soort activiteiten. Dit roept wel privacy zorgen op, omdat het privé-informatie in de chat analyseert. Waardoor de privacy van de gebruiker bij geheime chatberichten worden aangetast.

Do you think the technology invades the privacy of the stakeholders? If yes, in what way?

This question has not been answered yet.

Is the technology is compliant with prevailing privacy and data protection law? Can you indicate why?

This question has not been answered yet.

Does the technology mitigate privacy and data protection risks/ concerns (privacy by design)? Please indicate how.

This question has not been answered yet.

In which way can you imagine a future impact of the collection of personal data?

This question has not been answered yet.

Now that you have thought hard about privacy and data protection, what improvements would you like to make? List them below.

Om privacy en gegevens te beschermen van de gebruikers is het vooral van belang om te anonimiseren, hierdoor hou je de gegevens veilig bij het scannen van berichten. Zonder dat je inbreuk doet op persoonlijke gegevens. De algoritmes moeten ook transparant werken, gebruikers moeten begrijpen hoe de algoritme werkt en waarvoor het wordt gebruikt. Het is ook belangrijk dat er een duidelijke toestemming knop komt, voordat deze functie wordt ingeschakeld.

Human values

How does the technology affect your human values?

This category is only partial filled.

How is the identity of the (intended) users affected by the technology?

De identiteit van mogelijke gebruikers kan worden beïnvloed doordat hun privacy mogelijk wordt geschaad door AI-contentanalyse. De gebruikers kunnen zich beperkt voelen met hun zelfexpressie op het platform, door de monitoring. Verder kan het angst opwekken, doordat de anonimiteit wordt verminderd. Het kan ook positief ervoor zorgen dat bepaalde gebruikers zich minder zorgen hoeven te maken over het ontvangen van schadelijke of haatdragende berichten.

How does the technology influence the users' autonomy?

This question has not been answered yet.

What is the effect of the technology on the health and/or well-being of users?

This question has not been answered yet.

Now that you have thought hard about the impact of your technology on human values, what improvements would you like to make to the technology? List them below.

Er moet meer transparantie zijn in hoe berichten of inhoud wordt gecensureerd, daarvoor moet de algoritme up-to-date gehouden worden. Dit kan gedaan worden door toevoeging van menselijke rapporten op bepaalde accounts of groepen, waardoor deze chats worden gemarkeerd door AI-contentanalyse. Het idee is om algoritmes te trainen tot eerlijk of onpartijdig, zodat het niet onterecht inhoud blokkeren of een groep onterecht uitsluit.

Stakeholders

Have you considered all stakeholders?

This category is only partial filled.

Who are the main users/targetgroups/stakeholders for this technology? Think about the intended context by answering these questions.

Name of the stakeholder
Sociale mediaplatformen

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Gebruikers

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Ai-ontwikkelaars

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Overheden

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Name of the stakeholder
Unknown

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Did you consider all stakeholders, even the ones that might not be a user or target group, but still might be of interest?
-

Now that you have thought hard about all stakeholders, what improvements would you like to make? List them below.
This question has not been answered yet.

Data

Is data in your technology properly used?

This category is only partial filled.

Are you familiar with the fundamental shortcomings and pitfalls of data and do you take this sufficiently into account in the technology?

-

How does the technology organize continuous improvement when it comes to the use of data?

This question has not been answered yet.

How will the technology keep the insights that it identifies with data sustainable over time?

This question has not been answered yet.

In what way do you consider the fact that data is collected from the users?

This question has not been answered yet.

Now that you have thought hard about the impact of data on this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Inclusivity

Is your technology fair for everyone?

This category is only partial filled.

Will everyone have access to the technology?
This question has not been answered yet.

Does this technology have a built-in bias?
-

Does this technology make automatic decisions and how do you account for them?
This question has not been answered yet.

Is everyone benefitting from the technology or only a a small group? Do you see this as a problem? Why/why not?
This question has not been answered yet.

Does the team that creates the technology represent the diversity of our society?
This question has not been answered yet.

Now that you have thought hard about the inclusivity of the technology, what improvements would you like to make? List them below.
This question has not been answered yet.

Transparency

Are you transparent about how your technology works?

This category is only partial filled.

Is it explained to the users/stakeholders how the technology works and how the business model works?

-

If the technology makes an (algorithmic) decision, is it explained to the users/stakeholders how the decision was reached?

This question has not been answered yet.

Is it possible to file a complaint or ask questions/get answers about this technology?

This question has not been answered yet.

Is the technology (company) clear about possible negative consequences or shortcomings of the technology?

This question has not been answered yet.

Now that you have thought hard about the transparency of this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Sustainability

Is your technology environmentally sustainable?

This category is only partial filled.

In what way is the direct and indirect energy use of this technology taken into account?

-

Do you think alternative materials could have been considered in the technology?

This question has not been answered yet.

Do you think the lifespan of the technology is realistic?

This question has not been answered yet.

What is the hidden impact of the technology in the whole chain?

This question has not been answered yet.

Now that you have thought hard about the sustainability of this technology, what improvements would you like to make? List them below.

This question has not been answered yet.

Future

Did you consider future impact?

This category is only partial filled.

What could possibly happen with this technology in the future?
In de toekomst kan dit AI-systeem steeds nauwkeuriger worden, dit kan leiden tot een betere detectie van schadelijke inhoudt in bepaalde groepen. Het kan er ook voor zorgen dat er een inbreuk komt bij de gebruikers van de privé chats, waardoor zij zich meer terughoudend opstellen.

Sketch a or some future scenario (s) (20-50 years up front) regarding the technology with the help of storytelling. Start with at least one utopian scenario.
This question has not been answered yet.

Sketch a or some future scenario (s) (20-50 years up front) regarding the technology with the help of storytelling. Start with at least one dystopian scenario.
This question has not been answered yet.

Would you like to live in one of this scenario's? Why? Why not?
This question has not been answered yet.

What happens if the technology (which you have thought of as ethically well-considered) is bought or taken over by another party?
This question has not been answered yet.

Impact Improvement: Now that you have thought hard about the future impact of the technology, what improvements would you like to make? List them below.
De volgende verbetering zou ik graag uitvoeren voor de technologie:
- Gebruikers betere controle geven op de moderatie instelling van de AI-contentanalyse.
- Transparante algoritmes zodat gebruikers weten wat er met hun gegevens gebeurt.
- Onterechte censuur vermijden, door AI-systeem te trainen met juiste prompts.