Mitigating Child Trafficking on Meta's platforms

TKS Meta

Executive Summary

Problem

CHILD TRAFFICKING ON **META PLATFORMS**

20M instances of shared **sexual abuse material** on Meta platforms.

Meta is unable to detect trafficking on DMs- it cannot contact police or parents.

Solution

SCANNING **DMs** FOR **MALICIOUS INTENTION** & **INFORMING** CHILDREN

Take **proactive measures** to prevent the child from getting attached to the trafficker.

Make them **reflect on the situation** & realize the risk they're in.

Outcomes

PREVENTING CHILDREN FROM BEING **HARMED** IN ANY CASE

Notify parents about threats to children's safety online.

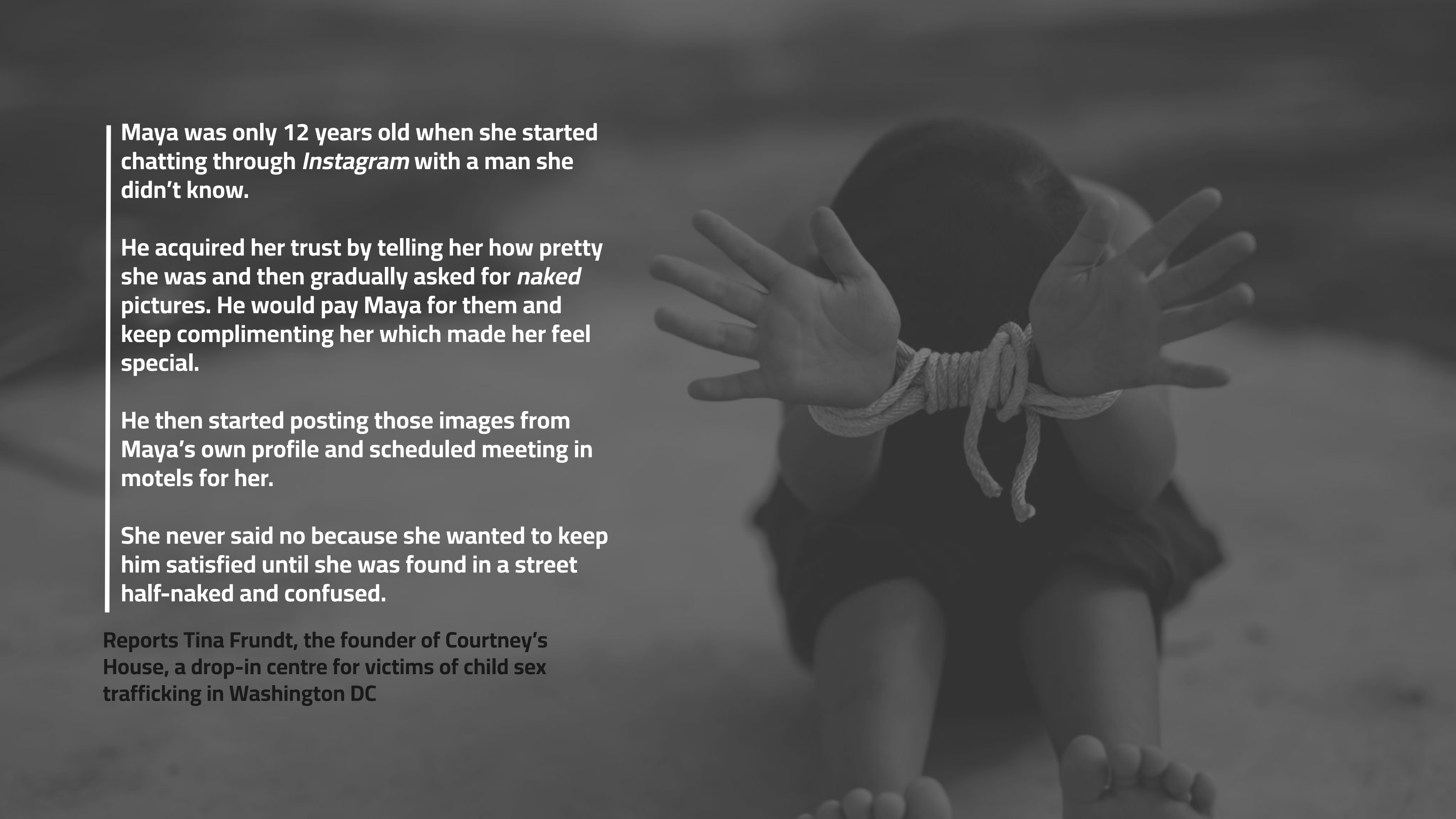
Educating **children to stay safe**, while accessing Meta platforms.

Human Trafficking is the **3rd** largest crime sector after drugs and guns

Human trafficking brings **150 billion** in business annually.

10.1 million **human trafficking** victims
are under 18.

65% of victims are physically abused, of sexually assaulted



> Maya was only 12 years old when she started chatting through *Instagram* with a man she didn't know.
>
> He acquired her trust by telling her how pretty she was and then gradually asked for *naked* pictures. He would pay Maya for them and keep complimenting her which made her feel special.
>
> He then started posting those images from Maya's own profile and scheduled meeting in motels for her.
>
> She never said no because she wanted to keep him satisfied until she was found in a street half-naked and confused.

Reports Tina Frundt, the founder of Courtney's House, a drop-in centre for victims of child sex trafficking in Washington DC

01

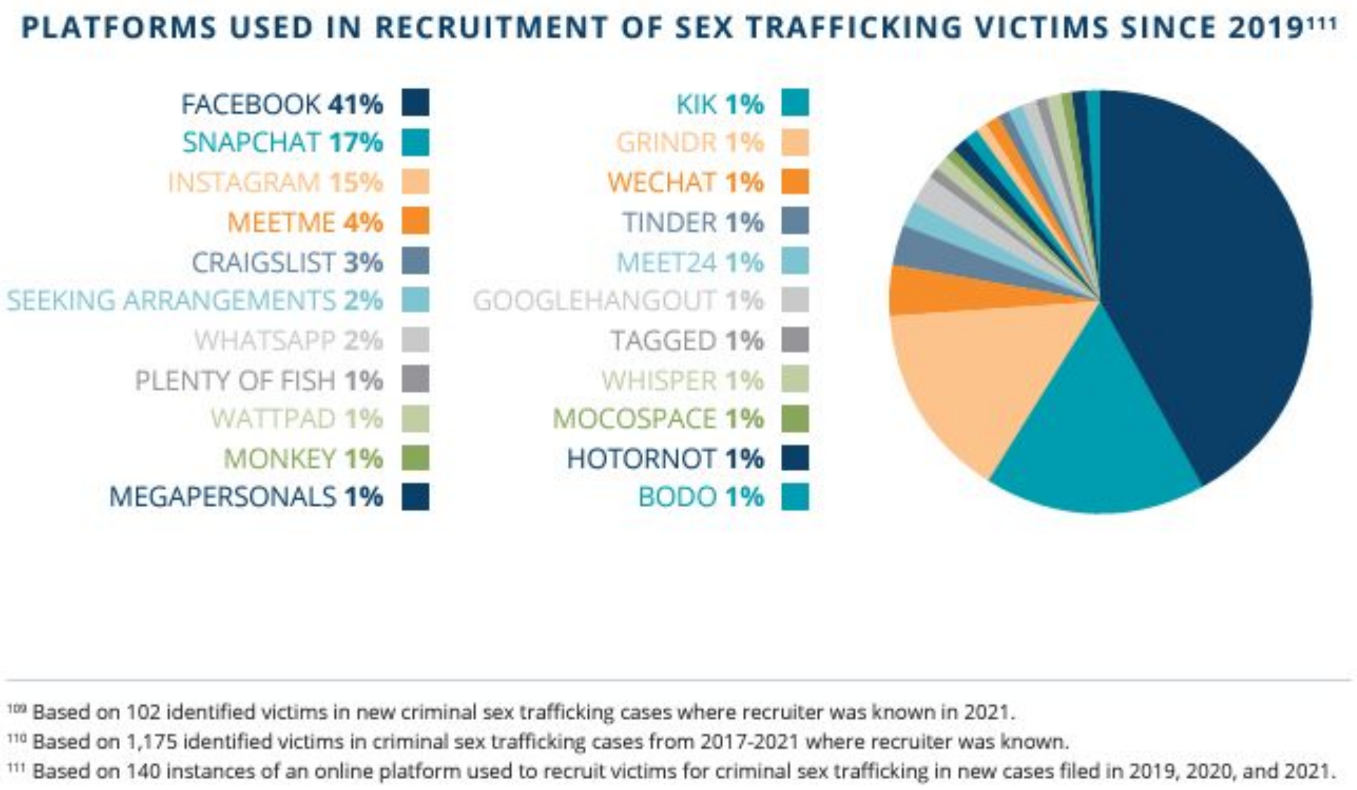
Problem

Why is social media used by traffickers?

366 federal criminal cases in the U.S. that featured suspects using Facebook for child exploitation

Social media has been used by traffickers to recruit victims, to spread their operations, and to have control over victims through limiting their ability to access their social media accounts by impersonating the victim, or spreading lies and rumors online.

The process of trafficking usually starts with the trafficker and the potential victim building a trusting relationship through social media.



PLATFORMS USED IN RECRUITMENT OF SEX TRAFFICKING VICTIMS SINCE 2019[111]

FACEBOOK	**41%**	KIK	**1%**
SNAPCHAT	**17%**	GRINDR	**1%**
INSTAGRAM	**15%**	WECHAT	**1%**
MEETME	**4%**	TINDER	**1%**
CRAIGSLIST	**3%**	MEET24	**1%**
SEEKING ARRANGEMENTS	**2%**	GOOGLEHANGOUT	**1%**
WHATSAPP	**2%**	TAGGED	**1%**
PLENTY OF FISH	**1%**	WHISPER	**1%**
WATTPAD	**1%**	MOCOSPACE	**1%**
MONKEY	**1%**	HOTORNOT	**1%**
MEGAPERSONALS	**1%**	BODO	**1%**

[109] Based on 102 identified victims in new criminal sex trafficking cases where recruiter was known in 2021.

[110] Based on 1,175 identified victims in criminal sex trafficking cases from 2017-2021 where recruiter was known.

[111] Based on 140 instances of an online platform used to recruit victims for criminal sex trafficking in new cases filed in 2019, 2020, and 2021.

Based on the 2021 Federal Human Trafficking Report

Human Trafficking Victim Demographics

Particularly vulnerable youth populations

Certain young people are more vulnerable to being trafficked than others, and there are often factors that amplify the chances of becoming potential victims.

Particularly vulnerable groups of teenagers share backgrounds of poverty, family issues physical and sexual abuse, and trauma.

Racial and ethnic minority teenagers are more vulnerable to trafficking because they're connected to poverty.

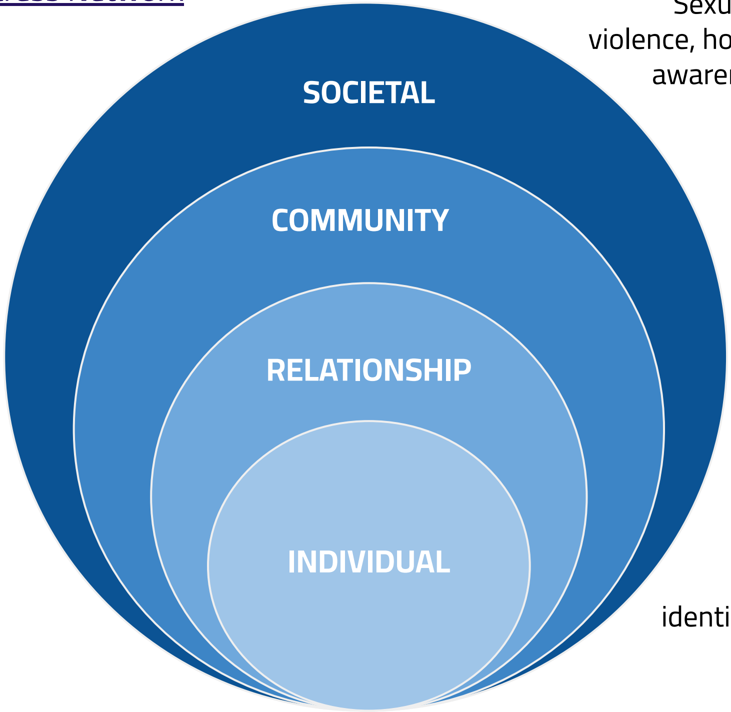
Rather than these general demographic groups, certain populations of youth are at high risk for being trafficked.

Factors influencing vulnerability

Based on the Child Traumatic Stress Network

Sex trafficking occurs among all socioeconomic classes, races, ethnicities, and gender identities.

However, some youth are at increased risk due to a complex interaction of **societal, community, relationship, and individual factors.**



Sexualization of children, gender-based violence, homophobia and transphobia, lack of awareness and resources, social injustice

Under-resourced schools and neighborhoods, community social norms, gang presence, commercial sex in the area, poverty and lack of employment opportunities.

Family dysfunction, intimate partner violence, caregiver loss or separation

Abuse/neglect, systems involvement (child protection, juvenile justice), runaway, LGBTQ identity, intellectual and/or developmental disability, mental health concerns, substance use, unaccompanied migration status

60% of child sex trafficking victims are or have been in foster care

How do traffickers find their victims in social media?

If a young person is lonely, they might talk about it on social media or participate in chat rooms looking for understanding.

These searches might feel innocent but they represent the emptiness individuals are trying to fill.

Human traffickers are used to *identifying* these markers and using the right words to taking advantage of these situations

Based on a study by the University of Toledo these are some examples of posts that might draw the attention of a trafficker as a sign of fear, emptiness and disappoint out of an individual's life

"Nobody gets me."

"I am so sick of being single."

"I am so ugly."

"How do I look?"

"My life sucks."

"She's not my true friend."

"My parents don't trust me."

"I'm being treated like a kid."

"I need to get out of here."

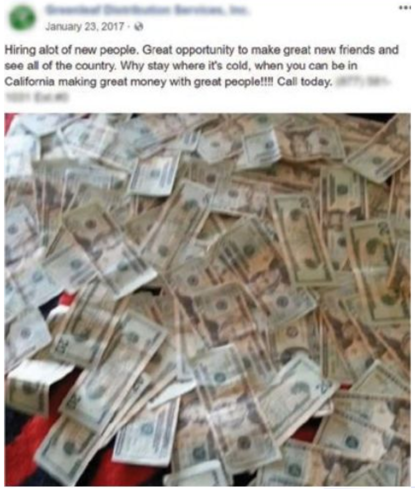
How do traffickers attract their victims?

Fake Business Profiles

Certain sex traffickers recruit victims through an illegitimate job offer, sometimes facilitated through fake business profiles or event pages

Misleading Photos

Traffickers often exaggerated images (such as stacks of cash), to lure individuals into clicking a link, and then providing their personal information.



[redacted]
January 23, 2017 ·

Hiring alot of new people. Great opportunity to make great new friends and see all of the country. Why stay where it's cold, when you can be in California making great money with great people!!!! Call today. [redacted]

Polaris report *"A Roadmap for Systems and Industries to Prevent and Disrupt Human Trafficking"*

Using DMs to communicate with the potential victim and build trust

Some individual sex traffickers may impersonate their bottom girl (a victim still under their control and valued higher than other victims).

A criminal may create fake social media accounts specifically to interact with school-aged children in this process. They may use the fake profiles to chat online for months, pretending to have common interests and **building trust.**

Traffickers may also contact a potential victim directly, claiming to be a recruiter for a legitimate business seeking staff. They may recruit victims abroad with a “scholarship” to a U.S. university.

The process followed in DMs

Stage 1:

Recruitment through social media may begin with commenting on potential victims' photos and sending direct messages by carefully building trust and intimacy

Stage 2:

The next phase is the so called "boy-friending" – manipulations such as feigned romantic interests, extreme flattery, promises of gifts or relevan financial assistance, assurance that they care for the potential victim, or even perceived savery from domestic violence or child sexual abuse

Stage 3

In such cases, the online relationship will generally move forward with the trafficker purchasing travel tickets for the potential victim in order to finally unite face-to-face. After the first meeting, the grooming stage takes place.

⚠ The final stage of the process is trafficking and **control**. The trafficker uses violent threats, withholds necessary resources, and engages in explicit acts of abuse to keep the victims in their control.

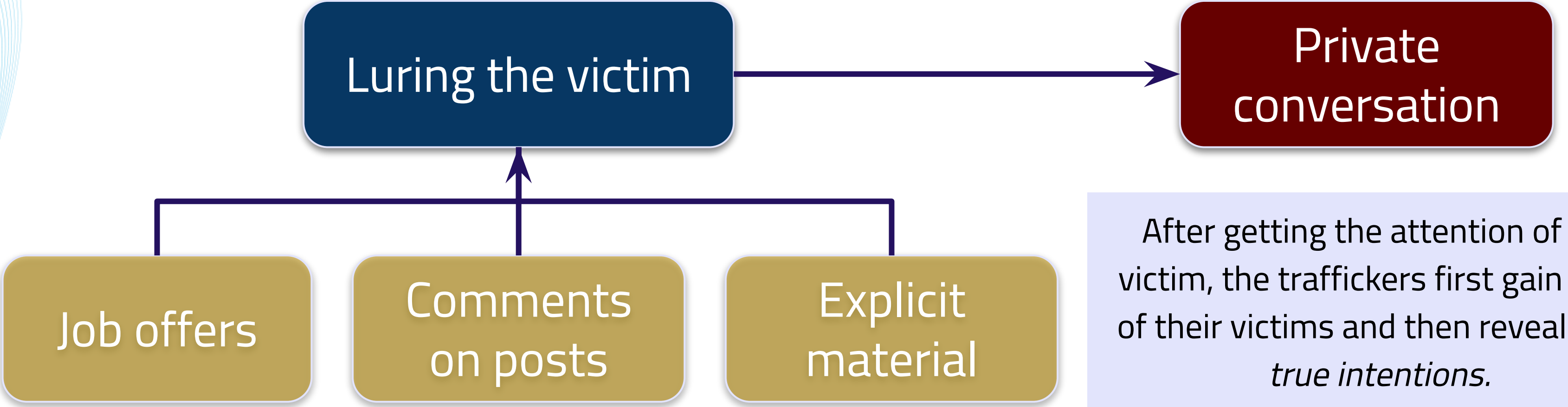
How has Meta acted on these issues?

- ✓ Restricting adult -> teen DMs without following
- ✓ Hiding suspicious adults from recommendations & search
- ✓ Provided tools to report the trafficker

- ✗ Contact can be gained before DMs
- ✗ Inefficient filtering of such accounts from search
- ✗ Rarely affecting children's decision due to their trust in the trafficker

More detailed approach in Appendix

Where is the gap? - Private Messages



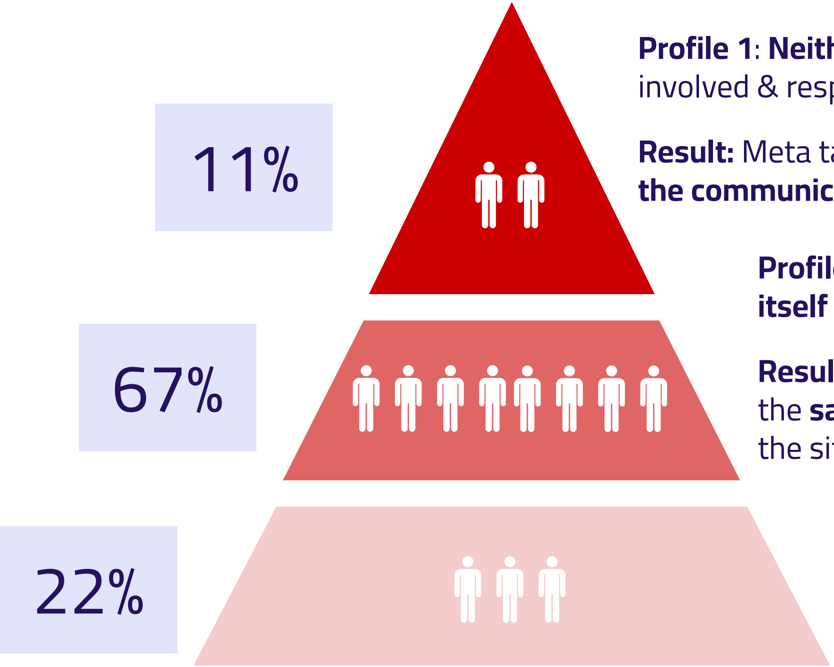
After getting the attention of the victim, the traffickers first gain trust of their victims and then reveal their *true intentions.*

Nothing is being done to check the intentions behind the DMs of innocent children that eventually fall victim to child trafficking.

02

Solution

Our solution addresses different victim profiles



Profile 1: **Neither the family environment nor the child** are involved & responsive to Meta's warnings.

Result: Meta takes **proactive measures** by entirely **blocking the communication** between the trafficker and the victim

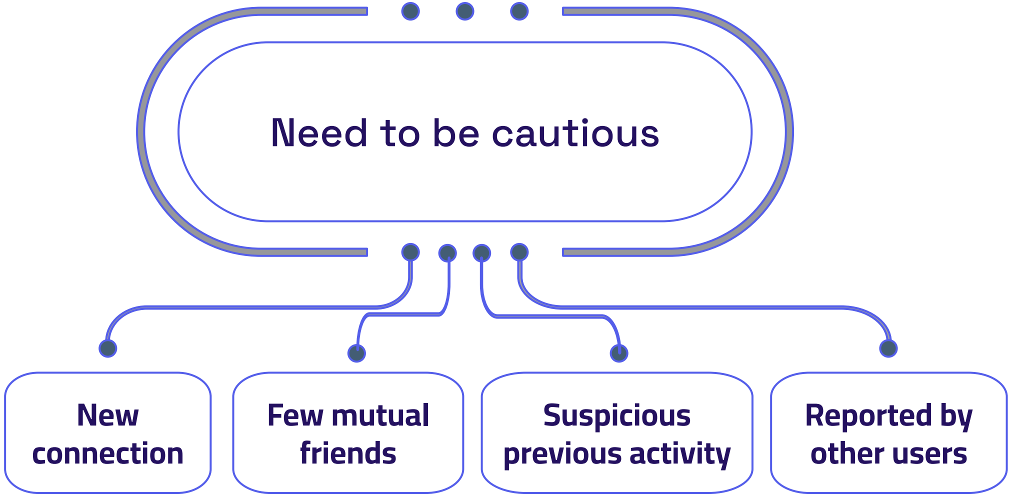
Profile 2: Child is in **abusive or careless** environment, but is **itself** aware of the dangers

Result: Analysis of possible malicious intentions, along with the **safety questions**, for their **education** & **realization** of the situation, to **protect** themselves

Profile 3: Children in a **supportive & cautious family environment**

Result: Parents **taking action after getting the report** of suspicious communication with their child.

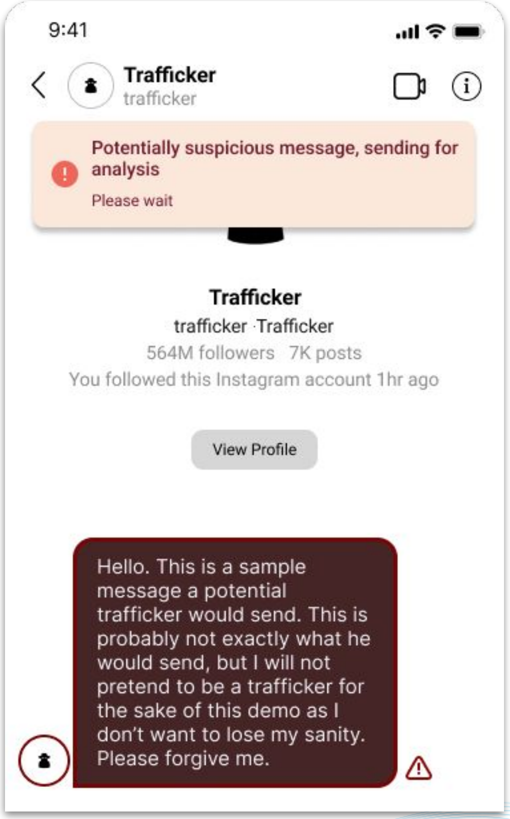
Supervision Measures



1 Detecting potentially suspicious account

We check if the account who sent a message could be a potential trafficker, by checking if it’s a new connection, has very few mutual friends with the child, and has potentially suspicious previous activity or has been reported before.

Supervision Measures



9:41

Trafficker
trafficker

Potentially suspicious message, sending for analysis

Please wait

Trafficker

trafficker ·Trafficker

564M followers 7K posts

You followed this Instagram account 1hr ago

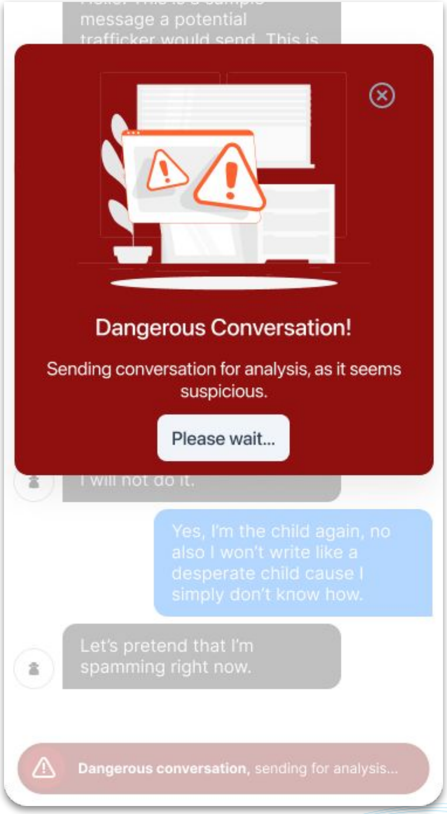
View Profile

Hello. This is a sample message a potential trafficker would send. This is probably not exactly what he would send, but I will not pretend to be a trafficker for the sake of this demo as I don't want to lose my sanity. Please forgive me.

2 Analyzing initial message's intention

Send the first message from the child's device to Meta's servers, where a Large Language Model, able to predict the intention of the message and provide relevant reasons in a concise way, will analyze it.

Supervision Measures



message a potential trafficker would send. This is

Dangerous Conversation!

Sending conversation for analysis, as it seems suspicious.

I will not do it.

Yes, I'm the child again, no also I won't write like a desperate child cause I simply don't know how.

Let's pretend that I'm spamming right now.

Dangerous conversation, sending for analysis...

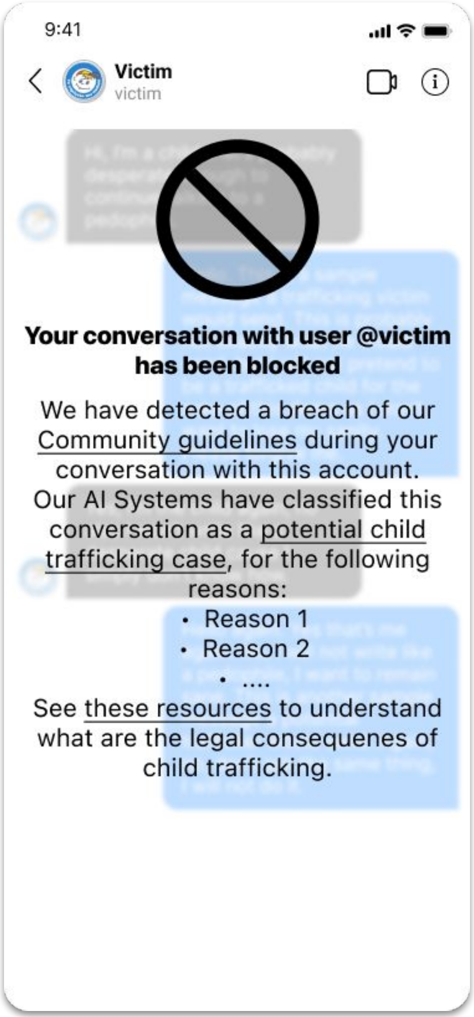
3 Supervising the conversation over time

5 days after the initial connection, and when a suspicious keyword is detected, send the messages to Meta's servers and use the same LLM as above to predict the intention of the entire conversation, or specific segments of it, providing relevant reasons.

Actions & restrictive Measures

4 Block conversation & interactions

If the contact seems suspicious, even after an initial warning from Meta (*2 strikes*), block Direct Messaging between those two accounts, and every interaction between them (comments, likes, recommend etc.). Inform to the kid about the apparent risk with a concise analysis from the LLM.

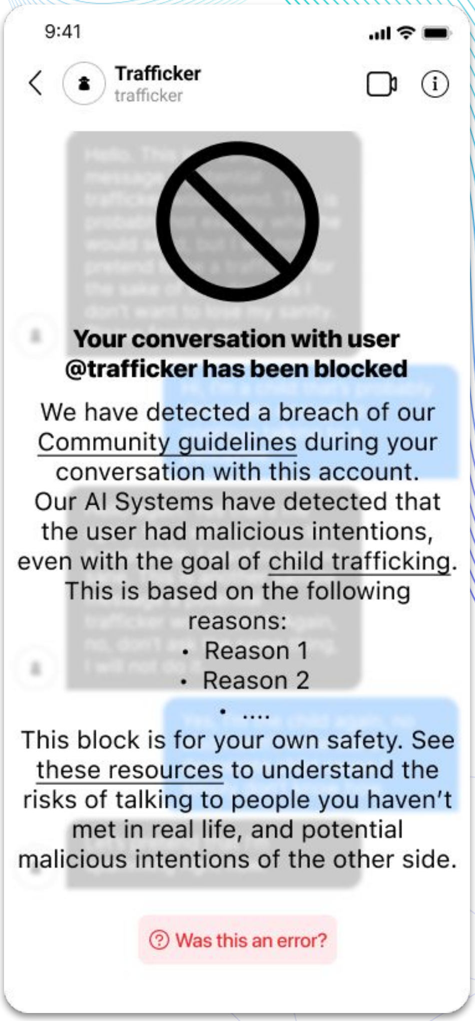


Your conversation with user @victim has been blocked

We have detected a breach of our Community guidelines during your conversation with this account. Our AI Systems have classified this conversation as a potential child trafficking case, for the following reasons:

- Reason 1
- Reason 2
- ....

See these resources to understand what are the legal consequenes of child trafficking.



Your conversation with user @trafficker has been blocked

We have detected a breach of our Community guidelines during your conversation with this account. Our AI Systems have detected that the user had malicious intentions, even with the goal of child trafficking. This is based on the following reasons:

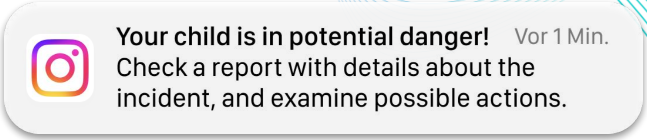
- Reason 1
- Reason 2
- ....

This block is for your own safety. See these resources to understand the risks of talking to people you haven't met in real life, and potential malicious intentions of the other side.

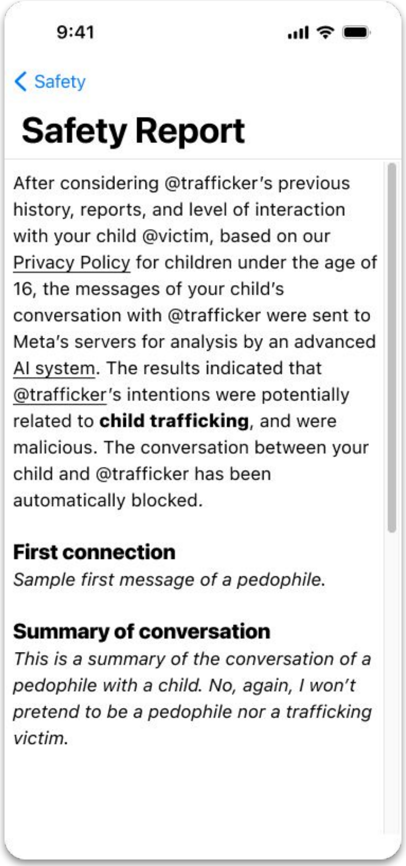
Actions & restrictive Measures

5 Inform the parent

If there is a registered parent for supervision, send them the first message & a summary of the conversation, along with an analysis of the LLM with the reasons the suspicions were raised.



Your child is in potential danger! Vor 1 Min.
Check a report with details about the incident, and examine possible actions.



Safety Report

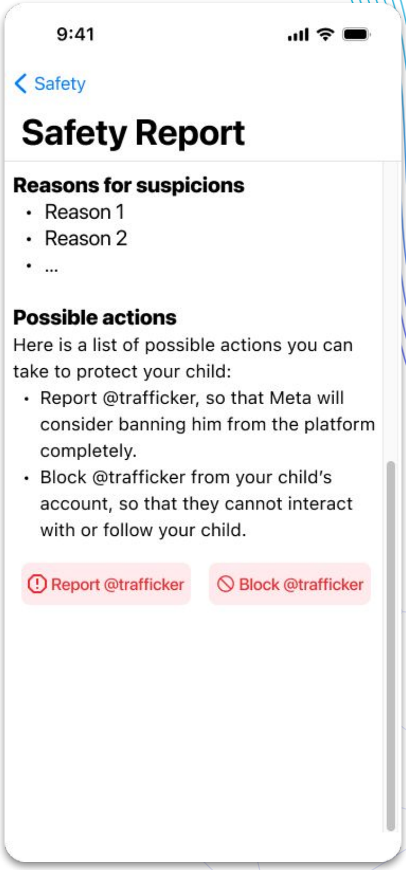
After considering @trafficker's previous history, reports, and level of interaction with your child @victim, based on our Privacy Policy for children under the age of 16, the messages of your child's conversation with @trafficker were sent to Meta's servers for analysis by an advanced AI system. The results indicated that @trafficker's intentions were potentially related to **child trafficking**, and were malicious. The conversation between your child and @trafficker has been automatically blocked.

First connection

Sample first message of a pedophile.

Summary of conversation

This is a summary of the conversation of a pedophile with a child. No, again, I won't pretend to be a pedophile nor a trafficking victim.



Safety Report

Reasons for suspicions

- Reason 1
- Reason 2
- ...

Possible actions

Here is a list of possible actions you can take to protect your child:

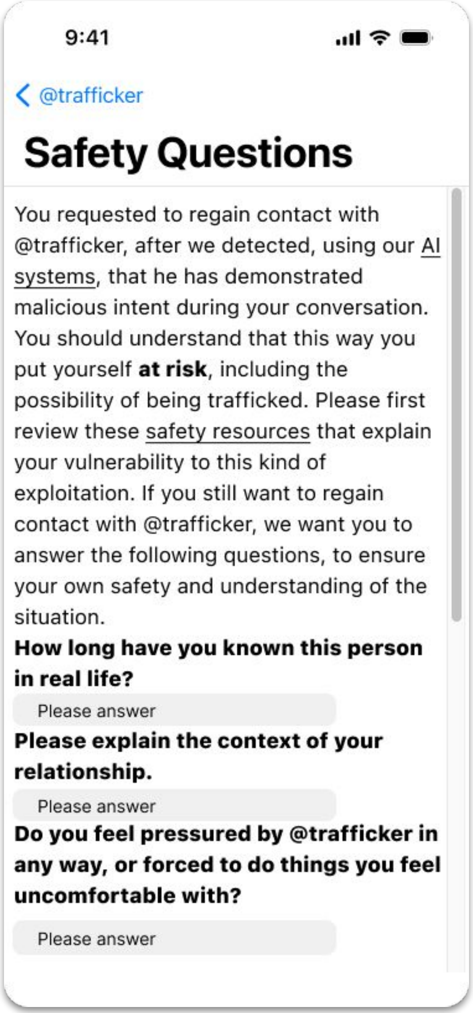
- Report @trafficker, so that Meta will consider banning him from the platform completely.
- Block @trafficker from your child's account, so that they cannot interact with or follow your child.

Actions & restrictive Measures

6 Safely restoring contact upon request

If the kid wants to regain contact with the person (possibly due to a false positive of the algorithm), then send the messages to a human reviewer, or request parental consent.

After success of either option, ask the kid some safety questions, personalized using an LLM and based on the conversation with the trafficker, with the purpose of it realizing the intention & possible risks.



< @trafficker

Safety Questions

You requested to regain contact with @trafficker, after we detected, using our AI systems, that he has demonstrated malicious intent during your conversation. You should understand that this way you put yourself **at risk**, including the possibility of being trafficked. Please first review these safety resources that explain your vulnerability to this kind of exploitation. If you still want to regain contact with @trafficker, we want you to answer the following questions, to ensure your own safety and understanding of the situation.

How long have you known this person in real life?

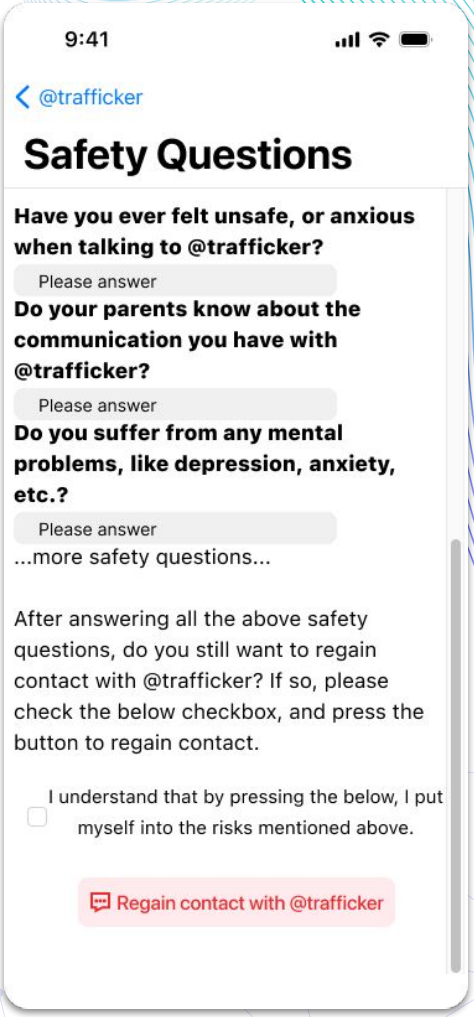
Please answer

Please explain the context of your relationship.

Please answer

Do you feel pressured by @trafficker in any way, or forced to do things you feel uncomfortable with?

Please answer



< @trafficker

Safety Questions

Have you ever felt unsafe, or anxious when talking to @trafficker?

Please answer

Do your parents know about the communication you have with @trafficker?

Please answer

Do you suffer from any mental problems, like depression, anxiety, etc.?

Please answer

...more safety questions...

After answering all the above safety questions, do you still want to regain contact with @trafficker? If so, please check the below checkbox, and press the button to regain contact.

□ I understand that by pressing the below, I put myself into the risks mentioned above.

Regain contact with @trafficker

Privacy Policy

For children under the age of 16, Meta will have **access to any DM** sent to the child from an account that has been classified as **suspicious**, based on **publicly available indicators**.

The DMs that Meta will have access to, will only be sent to the servers in a situation with possible **imminent danger**, where the messages will be processed by an algorithm or a human, but **deleted right after processing**.

Not storing the data in a non-encrypted format for a prolonged period of times **eliminates the possibility of a personal data breach**

Given that the **child is under the supervision of a parent** registered in Meta's platforms, their DMs that are processed by the Machine Learning algorithm may be converted to a **report** and **sent to the parent**.

Why Will This Work?

Builds upon the idea of conscious interruptions

This elevates Instagram's initiative of safety warnings to more meaningful heights by targeting the audience autonomously.

The solution becomes more *Proactive* than *Reactive*

The depth of the solution minimizes the reliability on the awareness of the users and *instead* educates them on how they might be in danger.

Provides an immediate solution everywhere, everytime

It ensures that all conversations flagged as suspicious are monitored carefully in order to provide any assistance at any point in time.

Mental & educational impact to the child

Awareness:
Informing the child about the potential dangers of the situation they're in by reasoning why the other side may have **malicious intentions**.

Reflection:
If the child wishes to regain contact with the other side, safety questions make it **reflect on the situation** & understand the **threat** posed by the other side.

Education:
By **educating the child** about the risks, we equip them with the **critical thinking and knowledge** necessary to **get out** of the possibly vulnerable situation.

Social media trafficking cases stopped

15% Of children that had a potentially harmful online experience **ignored it**

51% Thought it wasn't a big deal

21% Of them were worried they would get in trouble with their family

46% Have been contacted using a private message service

37% Have been contacted by the same user, on the same platform, under a new identity

We'll save all of them. It **is** a big deal, and that's why we'll **stop** it in most cases.

Based on Thorn. (2023). Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking in 2021.

Facilitating parents’ actions

67% Of parents felt **they** were they ones responsible for their children’s safety

55% Of those who knew, deleted/blocked the suspicious person contacting their child

But **71%** of them **don’t know**

46% Of children don’t realize they are in danger.

44% Are scared of the parents’ reaction.

We’ll **inform the parents** in any case, giving them the **power** to save their child.

Based on Thorn. (2023). Responding to Online Threats: Minors’ Perspectives on Disclosing, Reporting, and Blocking in 2021.

Informing children about risks & understanding their mental state

67%	Of minors want more information on how to protect themselves.
93%	Struggled with drug abuse/mental health issues.
96%	Experienced physical/sexual/ emotional abuse.

We'll provide them with resources on **protecting themselves** on the internet & understanding potentially harmful situations.

We'll also help them **reflect on their mental state** and understand their **real reasons of connection** with a malicious person, making them realize the harm this may cause them.

Based on "Survey Results from the First National Survivor Study"

How could we implement this?

Fine-tuning LLaMA - 2 as an identification method of suspicious text

Repurposing LLaMA - 2 for this task not only saves resources required to implement the solution but also the time taken. Plus the 65 Billion parameters that the LLM is trained upon adds another layer of accuracy and credibility.

Human intervention in times of false positives

There may be cases of inaccurate judgement of the conversation by the LLM and the users can contest that decision and the appeal would then be forwarded to a human reviewer to check the legitimacy of the action made by the LLM.

The Timeline for building the LLM

Stage 1:

Gathering data for the training of the LLM. The data can come from Meta's database of reported child trafficking cases as well as a vocabulary bank of suspicious words or wordings. Data would then have to be labelled appropriately.

Stage 2:

The data would then be fed into the foundational model of LLaMA - 2 where – based on the amount of data available – and parameters would be adjusted to give the identification model with an accuracy figure.

Stage 3

The model would then be trained on the remaining dataset until a particular accuracy quota has been met. Further human testing would then be conducted before being released.

⚠ Based on the level of accuracy of the model and the result it produces, an action out of the security measures introduced earlier would take place.

The 4 Stages of Suspicion.

0-30%

Asking the user if the conversation feels suspicious.

30-50%

Providing an option to forward the profile to parent/guardian.

50-80%

Temporarily blocking the users’ contact with the person and educating them about safety hazards.

80-100%

Permanently blocking user contact and reporting profile to local authorities.

Limitations

- There may be **false positives**, blocking non-harmful conversations and worsening children's experience, but we choose the safer side.

- The LLM won't be completely **accurate** initially, as it will improve using real data. It would, however, adapt on user feedback.

- New types of covert communication and **hidden messages** could possibly slip through the algorithm and not get blocked.

Which people will be responsible for implementing this?

Meta’s Child Safety Task Force

Meta’s Fundamental AI Research Team

Financial Analysis

Assumptions made:

- 65 Billion Parameters
- Similar training efficiency as popular LLMs
- Using NVIDIA A100 GPUs for processing ($72 per day)
- 100,000 GPU days required for complete training
- 200,000 tokens per second for one A100 GPU
- 12.5 DMs per day
- 125 characters per DM
- 5 parallel GPU's (threading)

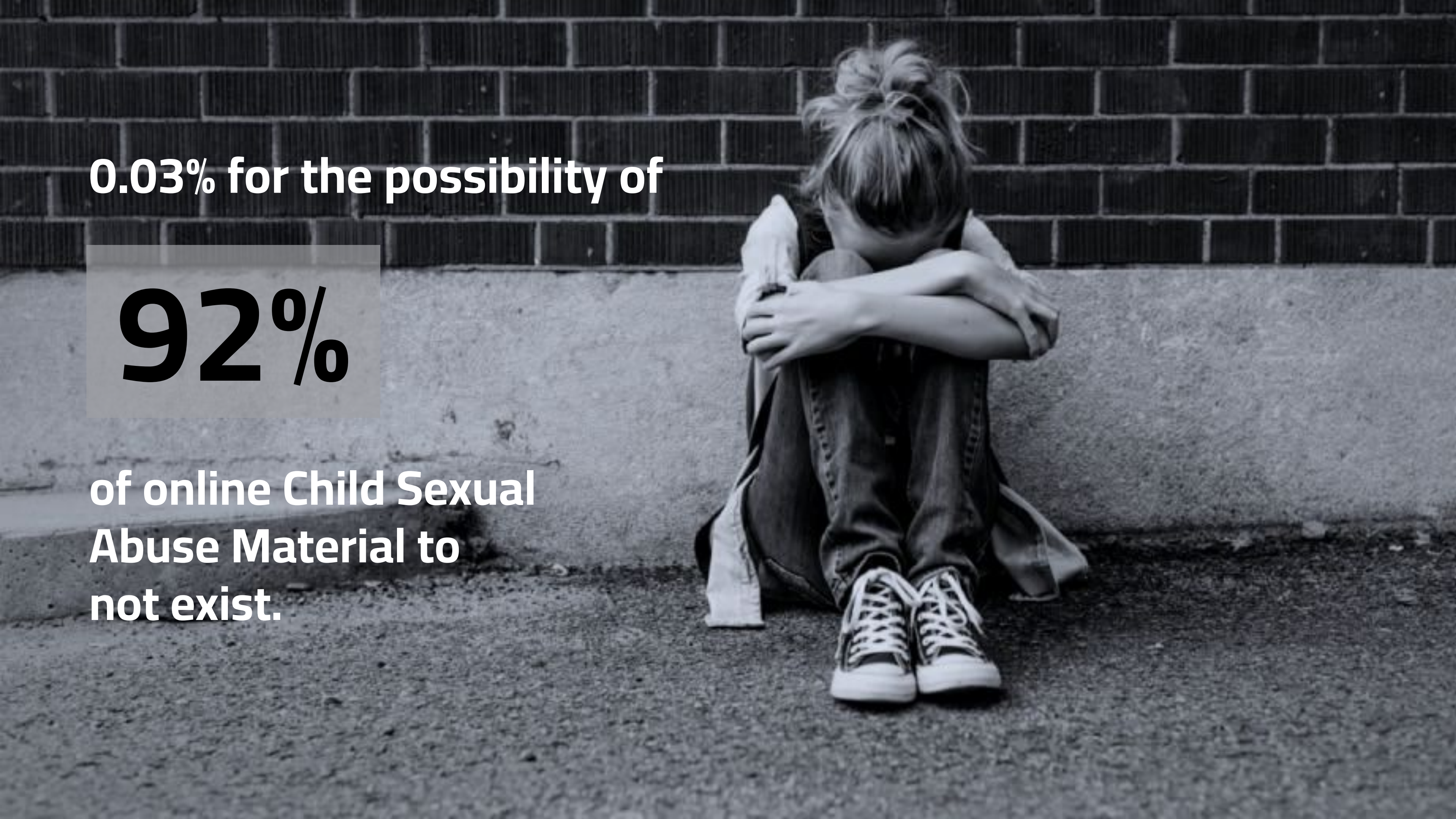
Calculations:

- 100,000 * 72 = $7.2 Million dollars – Total computing cost
- Daily processing of (1.5625 Billion tokens / 200,000) * 5 = 10.85 GPU hours daily
- Daily processing cost = 10.85 * 3 * 5 = $ 162.75
- Monthly cost = 163.75 * 30 = $ 4,882.5

Total cost adds up to $7.2 Million dollars plus $4,882.50 monthly expenses which would be

0.03%

Of the annual operating expenses for Meta softwares (In 2023).



0.03% for the possibility of

92%

of online Child Sexual Abuse Material to not exist.

03

The Metaverse

How will we implement our solution to the Metaverse?

- Instead of only checking DMs, we will translate voice to text, when the user is inside a virtual space.
- We can incorporate body language signals into the Machine Learning detection algorithm.
- We will disallow inappropriate gestures and motions, or other signals, when using Augmented Reality.

Appendix

Meta’s failed attempts to mitigate trafficking

Restricting adults older than 19 years old from sending DMs to teens who don’t follow them, allowing a 2-year gap for peer connections.	Connection doesn’t start from DMs, but rather commenting on victim’s posts or the victim asking info about a “job opportunity”.
Not showing suspicious adults (reported or blocked before) in “People you may know” recommendations, and removing message button from them when viewing teen profiles.	“Suggested for you” algorithm suggests other accounts containing CSAM when viewing one with that content, and Search doesn’t effectively exclude such accounts from appearing.
Pop-ups restricting the search/sharing of CSAM, and AI algorithms filtering that material from appearing in suggestions.	The pop-ups are based on detecting a list of keywords in search query, which is inefficient, like the algorithms filtering CSAM, allowing lots of it.
Pop-ups appearing to children DMs, asking if they know the person in real-life, informing them about the risk of their conversation and the actions they can take. Also various resources about their online safety.	These pop-ups rarely affect the child’s decision to continue talking to a potential trafficker, while the informative resources are often ignored, also not being accessible enough.