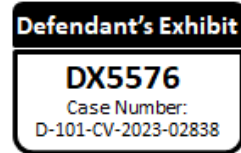


Message

From: Antigone Davis [/O=THEFACEBOOK/OU=EXTERNAL
(FYDIBOHF25SPDLT)/CN=RECIPIENTS/CN=████████████████████]
Sent: 1/15/2024 10:21:16 PM
To: Sadaf ██████ @meta.com]; Gabriel ██████ @meta.com]; Ravi Sinha ██████ @meta.com]; Antigone Davis ██████ @meta.com]; Nicole ██████ @meta.com]; Ashly ██████ @meta.com]; Kaitlin ██████ @meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":24230036373307842}]
Attachments: 418343930_902513424944812_1437162738539751231_n.png;
271547150_448982586770579_9096937701624976955_n.gif;
415838013_2263786707161621_4275552324638325217_n.png;
414027765_681266184089948_5351516231531587778_n.png;
418313911_1203219184398973_4293067319902408064_n.png;
418313911_1203219184398973_4293067319902408064_n.png



Gabriel ██████ (1/15/2024 10:40:26 PST):
>Hey!
>
>(1) Yes on PYMK, but with the specifics of
>a/ MaCSA>0.3 to Minor: Filtered
>b/ Minor to MaCSA >0.3: Filtered except where 2 year age gap
>c/ Minor to MaCSA >0.6: Filtered with no age gap
>
>(2) There were some filters in place prior to taskforce using UNIS criteria and MaCSA previously, but these were still leaving a large number of exposures in PYMK between MaCSA<>Minor so the Taskforce work was plugging the gaps:
>MaCSA Adult -> Minor <https://fb.workplace.com/groups/276817869676242/permalink/1163287271029293/>
>Bi-directional using UNIS criteria
<https://fb.workplace.com/groups/135709743752034/permalink/1011876329468700/>

Ravi Sinha (1/15/2024 10:40:48 PST):
>Thanks Gabe!

Ravi Sinha (1/15/2024 10:41:21 PST):
>Sadaf- Could we please consider revising the language in the deck?

Gabriel ██████ (1/15/2024 10:44:35 PST):
>+ kaitlyn, connecting the email thread from yesterday

Gabriel ██████ (1/15/2024 10:45:01 PST):
>+ kaitlin, connecting the thread from yesterday

Sadaf ██████ (1/15/2024 10:58:53 PST):
>hi hi, just want to make sure i understand
>
>"Minors are not recommended to potentially suspicious adults" was *not* in practice before the task force

Sadaf ██████ (1/15/2024 11:00:45 PST):
>im re reading it and the bullets are contradicting each other (unintentionally) a bit, right?
>
>the bullet says we don't rec minors to potentially suspicious adults before the task force

Sadaf ██████ (1/15/2024 11:01:09 PST):
>and then it later says "Meta had already implemented the second bullet: not allowing minors and potentially suspicious adults to be recommended to each other."

Sadaf ██████ (1/15/2024 11:01:22 PST):

shared: 418343930_902513424944812_1437162738539751231_n.png

Sadaf ██████ (1/15/2024 11:01:46 PST):
>is this a matter of fixing what the second bullet is referring to in the italicized bit?

Sadaf ██████ (1/15/2024 11:02:02 PST):
>or are both the 2nd and the 3rd bullet things that were implemented after the taskforce?

Sadaf ██████ (1/15/2024 11:05:41 PST):

PRIV

PRIV

Sadaf [REDACTED] (1/15/2024 11:05:59 PST):
>@Ravi @Kaitlin

Sadaf [REDACTED] (1/15/2024 11:18:58 PST):
>@Ashly [REDACTED]

PRIV

Kaitlin [REDACTED] (1/15/2024 11:19:09 PST):
>Sorry I'm a little lost on what we're trying to fix here (though noting that bullet three is redundant to bullet 1).

>
>Is the timing on what was in place before/after the Task Force a key distinction? It sounds like there was some filtering of not recommending minors to suspicious adults AND vice versa before the task force, but it was not comprehensive and the task force made it better. What's definitely wholly new as of the task force is not recommending suspicious adults to each other (or anyone).

Sadaf [REDACTED] (1/15/2024 11:20:12 PST):
>so the third bullet is net new

Sadaf [REDACTED] (1/15/2024 11:20:17 PST):
>Potentially suspicious adults are not recommended to each other.

Kaitlin [REDACTED] (1/15/2024 11:21:01 PST):
>And bullet 1 is net new (and by saying bullet 1 bullet three follows, so I'd put them together, like "Potentially suspicious adults are not recommended to anyone, including other potentially sus adults")

Sadaf [REDACTED] (1/15/2024 11:21:26 PST):
>the slide states, After the reporting was published, Meta's child safety task force reviewed its recommendation protections and implemented restrictions which informed the policies in place today, as discussed below.

>
>so it's ok that things are net new - the issue is the 4th bullet in italics says we already implemented the "2nd bullet" but then seemingly references the 3rd bullet

Sadaf [REDACTED] (1/15/2024 11:23:12 PST):
>so i guess its a matter of identifying if the 4th/background/italics bullet is in accurate?
>
>meaning before the task force, did we recommend minors and suspicious adults to each other?

Kaitlin [REDACTED] (1/15/2024 11:23:25 PST):
>Yeah we were already doing some of Bullet #2 prior to the task force, just poorly (Quoting from @Gabriel: "There were some filters in place prior to taskforce using UnIS criteria and MaCSA previously, but these were still leaving a large number of exposures in PYMK between MaCSA<>Minor so the Taskforce work was plugging the gaps"). I was not aware of us doing any of bullet 3 prior to the task force but I might be wrong there.}

Sadaf [REDACTED] (1/15/2024 11:24:11 PST):
>ok so adults not recommending to each other is net new.

Kaitlin [REDACTED] (1/15/2024 11:24:28 PST):
>My understanding - though I might be wrong - Gabe and Ravi are closer - is that we had no filtering of recommending potentially suspicious adults to each other prior to the TF. We DID have some (though inadequate) filtering of sus adults to minors and vice versa.

Sadaf [REDACTED] (1/15/2024 11:24:30 PST):
>minors were not recommended to adults prior to the task force, but the methods for that were improved during the task force

Kaitlin [REDACTED] (1/15/2024 11:24:44 PST):
>That's my understanding but @Gabriel [REDACTED] and/or @Ravi Sinha should confirm.

Sadaf [REDACTED] (1/15/2024 11:25:37 PST):
>and the issue is the language in the italics/4th bullet might be a bit murky bc it adds the 'recommended to each other' bit whereas the second bullet talks about just one way of recommendation -
>
>but we we didnt rec potentially suspicious adults to any users, that would include minors

Sadaf [REDACTED] (1/15/2024 11:26:28 PST):

shared: 271547150_448982586770579_9096937701624976955_n.gif

Kaitlin [REDACTED] (1/15/2024 11:28:28 PST):
>So there are four types of recommendations:
>1) Minors to Sus Adults
>2) Sus Adults to Minors

>3) Sus Adults to Sus Adults
>4) Sus Adults to ANYONE
>
>My understanding is both 1 & 2 existed before the TF but needed improvement and that 3 & 4 were net new. The current italics I believe also say that 1 & 2 existed prior to the TF, but it's written confusingly.

Sadaf [REDACTED] (1/15/2024 11:28:32 PST):
>thanks Kaitlin - Ravi - let me know what you think but if we were doin the 1st and 2nd bullet pre task force (albeit needing improvements) then the slide reads accurately to me

Ravi Sinha (1/15/2024 11:41:35 PST):
>I think we'll want to confirm that we were running both 1 and 2 on PYMK prior to the CSTF. My understanding--and I could be totally mistaken--is that our previous, IIC-based intervention was filtering minors out of IIC suspects' PYMK recommendations *but not the other way around.* Here's the test that I think we ended up building off of:

Ravi Sinha (1/15/2024 11:41:47 PST):

shared: 415838013_2263786707161621_4275552324638325217_n.png

Ravi Sinha (1/15/2024 11:41:52 PST):
><https://fb.workplace.com/groups/135709743752034/permalink/1011876329468700/>

Ravi Sinha (1/15/2024 11:42:45 PST):
>Here's more detail, that makes me think it was just unidirectional:

Ravi Sinha (1/15/2024 11:42:49 PST):
><https://docs.google.com/document/d/1-36TWHCZiBOW6GVZMfZSrw-Ln0JL7Nqf7usD-7ef4hA/edit#heading=h.wjihcttz38pk>

Ravi Sinha (1/15/2024 11:44:03 PST):
>And here's our canonical Child Safety interventions doc that Gabe and others own, which also makes me think that it was a unidirectional UNIS-based intervention on PYMK before the CSTF:

Ravi Sinha (1/15/2024 11:44:11 PST):

shared: 414027765_681266184089948_5351516231531587778_n.png

Kaitlin [REDACTED] (1/15/2024 11:44:20 PST):
>It sounds from Gabe like there was "bi-directional UNIS Criteria" pre-task force, so there for 1 & 2 but not all based on MacSA then.

Ravi Sinha (1/15/2024 11:44:20 PST):
>https://docs.google.com/document/d/1MVBd3zOQ3HhjVaYMsycw3DOaHFNzPKIzTzXKhePD_yQ/edit

Ravi Sinha (1/15/2024 11:45:41 PST):
>Honestly I may just be confused. Let's just confirm with Gabe and I'm happy to defer:
>
>@Gabriel- Was there a bi-directional IIC or UNIS based PYMK mitigation before the CSTF? Or was it only not recommending minors to IIC adults on PYMK before the CSTF?

Ravi Sinha (1/15/2024 11:46:48 PST):
>@Sadaf and @Kaitlin- If Gabe's answer is that it was that there was a bi-directional IIC or UNIS based PYMK mitigation before the CSTF, I'm of course happy with the language as drafted and hope that you'll forgive me for the holiday thrash. Really just not sure and want to make sure we're accurate here.

Sadaf [REDACTED] (1/15/2024 11:48:18 PST):
>we appreciate you - i also want to be super explicit since im not very used to these acronyms - the slide *currently says*
>
>Potentially suspicious adults are not recommended to any user *we DID do this before the task force*
>
>Minors are not recommended to potentially suspicious adults *we DID do this before the task force*
>
>- though this group is caveating to say it was improved after the task force

Kaitlin [REDACTED] (1/15/2024 11:50:03 PST):
>I am not aware of not recommending potentially sus adults to ANYONE as a thing we did before the task force (and sorry I don't know what full slide we're looking at so can't see what distinctions are being made on timing - I thought the list in the screenshot was all about what we did post TF)

Sadaf [REDACTED] (1/15/2024 11:50:27 PST):

shared: 418313911_1203219184398973_4293067319902408064_n.png

Sadaf [REDACTED] (1/15/2024 11:50:57 PST):
>(part of the slide not previously included in screenshot are just graphics) language from NM complaint and a WSJ article

Kaitlin [REDACTED] (1/15/2024 11:52:34 PST):

>Got it. I think my point still stands then. On what you said I think "Potentially suspicious adults are not recommended to any user we DID do this before the task force" is incorrect (at least as far as I'm aware) but I also don't think anything on the slide says this.

Sadaf [REDACTED] (1/15/2024 11:57:18 PST):

>ok ill wait for confirmation but keeping the slide as is for now and hoping it's accurate to say the first two bullets are accurate (thereby the slide is accurate though we understand there is additional details/context we may need to voice over if needed)

Antigone Davis (1/15/2024 12:02:46 PST):

>prior to task force it was limited to minors, post task force it was to anyone -- that is correct

Antigone Davis (1/15/2024 12:03:37 PST):

>if we need additional confirmation on all this Arcadiy is a source of truth

Sadaf [REDACTED] (1/15/2024 12:47:03 PST):

>hi Arcadiy - sorry for the cold add, want to get a quick gut check from you

Sadaf [REDACTED] (1/15/2024 12:53:51 PST):

>in this slide we say we don't recommend potentially suspicious adults to any users.

>

>we also say minors are not recommended to potentially suspicious adults (and that we did this in some part before the task force)

>

>*can you confirm that's accurate*

Sadaf [REDACTED] (1/15/2024 12:53:55 PST):

shared: 418313911_1203219184398973_4293067319902408064_n.png

Sadaf [REDACTED] (1/15/2024 12:58:52 PST):

>ah, apologies, Arcadiy is out on leave - @Ashly - as admin - can you remove him so he doesn't get bogged down with pings!

Antigone Davis (1/15/2024 13:13:23 PST):

>Sorry Arcadiy!

Sadaf [REDACTED] (1/15/2024 13:42:50 PST):

>ok just had a phone call with Ravi (thank you!)

>

>to be as clear as i can - the thing we *need to confirm is* -

>

>_Did we have restrictions in place, before the task force, on how/to whom we recommended suspicious adults? if so, what were those restrictions?_

>

>e.g., restricted to minors, certain groups of people, etc. If we can get to ground truth on that question, i think we'll be set.

>

>welcome getting final confirmation on that and any proof point that may exist to verify.

Gabriel [REDACTED] (1/15/2024 14:02:53 PST):

>So sorry, just back from a social event and missed this chat, just catching up now. Kate's message here is correct.

Sadaf [REDACTED] (1/15/2024 14:03:37 PST):

>don't be sorry! i know im just out here raining down pings at the worst times!

Sadaf [REDACTED] (1/15/2024 14:03:50 PST):

>just bc my brain is becoming slightly mushy, can i ask you to confirm this?

Sadaf [REDACTED] (1/15/2024 14:04:09 PST):

>in theory, it's saying the same thing as Kaitlin but my head has now read these slides 93849384 times

Gabriel [REDACTED] (1/15/2024 14:16:15 PST):

>>Did we have restrictions in place, before the task force, on how/to whom we recommended suspicious adults? if so, what were those restrictions?

>

>Yes. there was bi-directional risky adult <> minor filtering prior to the taskforce, though the interventions in each direction did not use the same filtering criteria.

>- Adults filtered from Minor PYMK recs using MacSA, migrated from previous IIC classifier. In the taskforce audit we found that this intervention was not filtering all recommendations and needed to re-implement it to ensure it was not leaking exposures.

>- Minor -> Adult using UnIS (legacy definition of risky pair). In the taskforce, we migrated this filter to the same MacSA threshold as used elsewhere, which added a much broader volume of exposures and required escalation with Growth. The outcome of this escalation was that we created a 2 year age gap carveout, but implemented the MacSA filtering for all other minor-> risky adult PYMK recommendations

>
>NET NEW delivered in taskforce
>- MaCSA 0.6 -> Filter from all users in PYMK
>- MaCSA 0.3 <> MaCSA 0.3 Filtering in PYMK

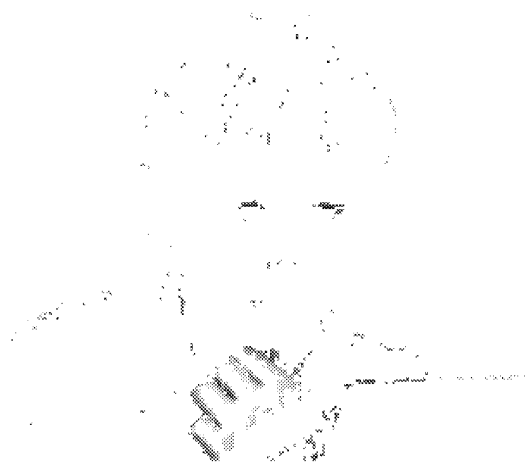
Sadaf [REDACTED] (1/15/2024 14:21:16 PST):
>ok got it - confirming both below applied before task force but were improved upon and net new is not
recommending suspicious adults to each other
>
>so potentially suspicious adults not recommended to users
>
>minors not recommended to potentially suspicious adults

Document Withheld for Privilege





























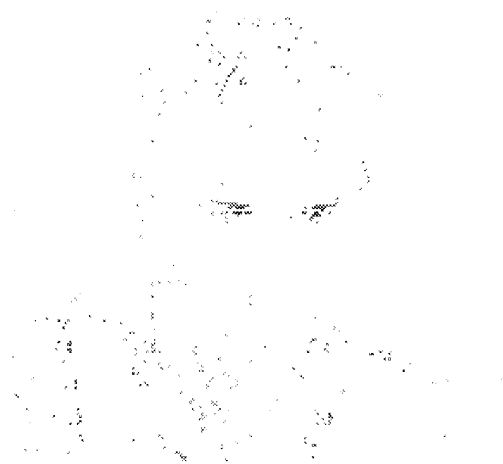
















































2	FB	Account recommendation filtering (via PYMK and Unconnected search) risky adults for / from minors	<p>Sept 2023 <u>PYMK filter</u> and <u>Search surface</u> launches:</p> <ol style="list-style-type: none"> 1. MaCSA 0.3 Adults to Minors 2. MaCSA 0.6 Adults to All 3. Minors to MaCSA 0.3 Adults (with 2 year age difference) <p>Previous:</p> <ol style="list-style-type: none"> 1. UnIS June 2022 <u>Launch</u> (global) <u>UnIS pair: FB V1 heuristic</u> (2022) Minor -> Adult filter 2. IIC <u>updated to MaCSA >0.3</u> Apr 2023 from previous IIC Classifier Adult -> Minor 	<p>MaCSA 0.3 Adult <> MaCSA 0.3 Adult (Sept 2023) <u>PYMK filter</u>, <u>Search surface launch</u>,</p>
---	----	--	--	---

Document Withheld for Privilege

Experiments ran for 6 weeks globally [Friending Circle review post]. The UnlS metric results exclude ePD threads (EU, Japan, UK etc.).

The treatment was:

- Control Group: no treatment will be applied.
- Test Group: victims will be filtered out from bad actor's PYMK suggestions.

Document Withheld for Privilege