MAY 24, 2021

BREAKING: Facebook Whistleblowers Expose LEAKED INTERNAL DOCS Detailing New Effort to Secretly Censor Vaccine Concerns on a Global Scale

Facebook has responded: "We proactively announced this policy on our company blog and also updated our help center with this information." - Facebook company spokesperson

Two Facebook Insiders have come forward with internal company documents detailing a plan to curb "vaccine hesitancy" (VH) on a global scale.

The stated goal of this feature is to "drastically reduce user exposure" to VH comments. Another aim of the program is to force a "decrease in other engagement of VH comments including create, likes, reports [and] replies."

It was such a shocking revelation, that it moved not just one -- but two whistleblowers to come forward to Project Veritas, so the public could be made aware of this plan to stifle free speech.

One Facebook whistleblower said the company uses a tier system to determine how a comment should be censored or buried.

Comments that include "shocking stories" describing potentially or actually true events, or facts that can raise safety concerns" -- are demoted.

"True events or facts" that raise concern about Covid vaccinations are fair game to be demoted and hidden -- according to our source -- despite their authenticity or capacity to contribute to the public good.

"I have to do something," one of the Facebook insiders said.

Project Veritas reached out to a top Facebook Spokesperson about these documents and received only a brief and broad statement in reply, that failed to address our biggest questions regarding transparency.

"They're trying to control this content before it even makes it onto your page before you even see it," the other Facebook insider added. "If I lose my job, it's like, what do I do? But that's less of a concern to me."

[Westchester, NY – May 24, 2021] Two Facebook Insiders have come forward to Project Veritas with leaked internal documents, showing the Big Tech giant's plan to police "Vaccine Hesitancy" (VH) through surreptitious "comment demotion."

The company has set up a tier system to rank comments on various scales, based on how much the statement questions or cautions against the Covid-19 vaccination.

Tier 2, for instance, represents "Indirect Discouragement" of getting vaccinated and according to PV's sources, user comments such as these would be heavily "suppressed."

It doesn't matter if the comments are true, factual or represent reality. The comment is demoted, buried and hidden from view of the public if it clashes with this system.

"It doesn't match the narrative," one insider explained. "The narrative being, get the vaccine, the vaccine is good for you. Everyone should get it. And if you don't, you will be singled out,"

One of the insiders, a Data Center Technician, leaked multiple internal documents detailing an algorithm test being run on 1.5 percent of Facebook and Instagram's nearly 3.8 billion users worldwide.

The goal? To, "drastically reduce user exposure to vaccine hesitancy (VH) in comments."

"They're trying to control this content before it even makes it onto your page before you even see it," one insider said.

1. The authors of the plan are credited as Joo Ho Yeo, Nick Gibian, Hendrick Townley, Amit Bahl and Matt Gilles.

Vaccine Hesitancy Comment Demotion

Credits

- Author List: @Joo Ho Yeo, @Nick Gibian, @Hendrick Townley, @Amit Bahl, @Matt Gilles
- Thanks:

Executive Summary

- What's your goal?
  - Drastically reduce user exposure to vaccine hesitancy (VH) in comments
- What is the product change?
  - Utilize the existing v1 VH classifier (English) to demote comments on ranked comments, meaning that they are filtered from 'most relevant' but are still visible in other tabs (ex 'most recent')
- What are the benefits of this launch?
  - VPVs on Vaccine post English comments vh p80: **-10.6±2.1%**
    - Projected launch impact: -934.8K±194.4K vpvs
    - Authoritative vh p80 comment vpvs: -26.7 (±4.1)%
      - Projected launch impact: -402.4K±71.3K
  - CEP on Vaccine post English comments vh p80: **–11.1 (±1.8)%**
    - Authoritative vh p80 comments CEP: **-26.1±3.0%**
  - Decrease in other engagement of VH comments including create, likes, reports, replies
    - ⚓Scuba grouped by VH
- What are the costs of this launch?
  - No significant cost is observed.
- Risks of this launch
  - Not all comments are actually vaccine hesitancy, but we'd aligned with Health Policy on this risk in the COVID Lockdown Decisions meeting 2 weeks ago – https://docs.google.com/presentation/d/1Qo35TGq75yf70-VkOAY61g0offjYaOB2o2dF6SUry44/edit#slide=id.gca2fb195a7_11_0
- How could this be made more aggressive?
  - Use lower thresholds for interventions
- How could this be made more conservative?
  - Use higher Thresholds for interventions

Background

Experiment Launch Post

Comments are a major surface relevant to our B2V efforts. We estimate that the prevalence of VH comments in Authoritative Health Pages is 25.3% and for other pages 19.42%. Now that the v1 Vaccine Hesitancy classifier has been cleared for this usecase, reducing the visibility of these comments represents another significant opportunity for us to remove barriers to vaccination that users on the platform may potentially encounter.

2. The goal of the "framework" is to "identify and tier the categories" of content that "could discourage vaccination in certain contexts."

Borderline Vaccine (BV) Framework

FKA: "Barriers to Vaccination (B2V), "Vaccine Hesitancy Content"

Goal: We aim to identify and tier the categories of non-violating content that could discourage vaccination in certain contexts, thereby contributing to vaccine hesitancy or refusal. We have tiered these by potential harm and how much context is required in order to evaluate harm.

Background: In Facebook Inc.'s COVID-19 vaccine Offense and Defense work, the goal is to understand how using the Family of Apps contributes to or detracts from vaccine uptake. The COVID-19 Health Behavior Change Framework illustrates how two types of drivers – drivers of intent and drivers of action – both contribute ultimately to vaccine behavior. To defensibly connect these categories of problematic vaccine-related content to potential harms, the framework below directly ties these types of non-violating content to the potential drivers of vaccination behavior. While these drivers relate to the larger idea of "vaccine hesitancy," this document more precisely and accurately describes factors that contribute to health behavior change. This Policy workstream reflects concerns that exposure to, interaction with, or production of that content can negatively impact these drivers (in other words, creating barriers to vaccination).

3. Facebook also established a "Vaccine Hesitancy" scoring system to establish numerical thresholds for certain content.

VH Classifier Score Threshold

	cleaned_score_bucket	adjusted_precision	adjusted_recall	precision_lower	precision_upper	recall_lower	recall_upper
0	0.0 - 0.3	0.126141	1.000000	0.085489	0.168945	1.000000	1.000000
1	0.3 - 0.5	0.126605	0.699953	0.084746	0.170880	0.554186	0.890621
2	0.5 - 0.6	0.197687	0.501959	0.147376	0.261657	0.381893	0.739008
3	0.6 - 0.65	0.386031	0.354294	0.330534	0.438976	0.264406	0.513907
4	0.65 - 0.7	0.462159	0.318081	0.395432	0.530412	0.234323	0.462183
5	0.7 - 0.75	0.490524	0.200457	0.426854	0.553555	0.143747	0.286414
6	0.75 - 0.8	0.588057	0.165619	0.514105	0.660887	0.119392	0.233647
7	0.8 - 0.85	0.677055	0.102807	0.598812	0.756986	0.072591	0.145819
8	0.85 - 0.9	0.730969	0.067448	0.648663	0.815614	0.046654	0.093516
9	0.9 - 0.95	0.744760	0.029917	0.656591	0.828318	0.020436	0.042050
10	0.95 - 1	0.795918	0.013449	0.680000	0.900000	0.009156	0.018791

Notebook: https://www.internalfb.com/intern/anp/view/?id=495202
Post: https://fb.workplace.com/notes/3942483092467702

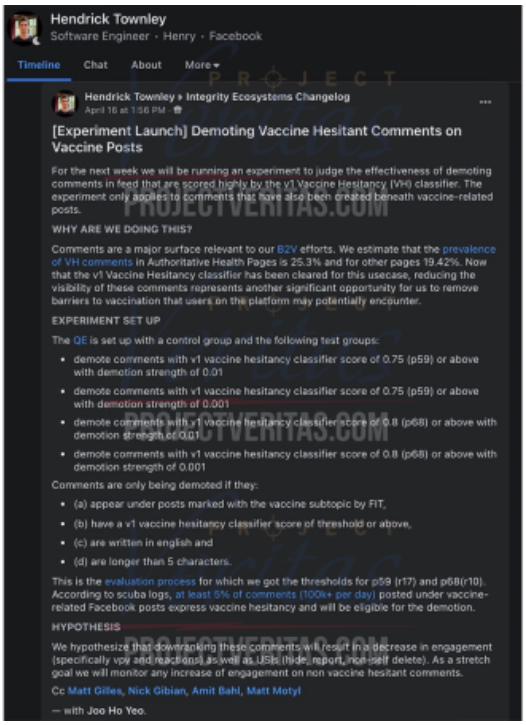
4. The company instituted tiers to describe the tone of certain posts, with some being categorized as "sensational/alarmist" or "criticizing choice." The responses range from outright removal to comment demotions.

Vaccine Hesitancy policy → action mapping

POLICY TIER	ACTION MAPPING
Violating: - COVID/vaccine M&H and WDH/RFH	**Remove** at scale
Violating: - Explicit Vaccine Discouragement - VH "sole purpose entities"	**Remove** on escalation
B2V Reduce *(fka Tiers 1/2)*: - Sensational/alarmist - Criticizing choice - Indirect discouragement (shocking stories; promoting alternatives)	*Everything from "Inform Only" tier, plus…* - **Non_rec** (inclusive of search, comments, and discovery surfaces) - **Demotions** (though not on indirect discouragement *content*) - **BTGs** if/as needed
B2V Inform Only *(fka Tiers 3/4)*: - Personal objections - Skepticism - Neutral discussion / debate	- **Inform** on *all* COVID Vaccine content + test more specific labels on top of that - **Reshare friction**

Policy note: For personal objections and skepticism, we have the ability to filter out comments if we want to.

5. Another document from the aforementioned Hendrick Townley discusses more data points following the "experiment" with "demoting" VH comments.



Hendrick Townley
Software Engineer · Henry · Facebook

Timeline Chat About More

Hendrick Townley › Integrity Ecosystems Changelog
April 16 at 1:56 PM

[Experiment Launch] Demoting Vaccine Hesitant Comments on Vaccine Posts

For the next week we will be running an experiment to judge the effectiveness of demoting comments in feed that are scored highly by the v1 Vaccine Hesitancy (VH) classifier. The experiment only applies to comments that have also been created beneath vaccine-related posts.

WHY ARE WE DOING THIS?

Comments are a major surface relevant to our B2V efforts. We estimate that the prevalence of VH comments in Authoritative Health Pages is 25.3% and for other pages 19.42%. Now that the v1 Vaccine Hesitancy classifier has been cleared for this usecase, reducing the visibility of these comments represents another significant opportunity for us to remove barriers to vaccination that users on the platform may potentially encounter.

EXPERIMENT SET UP

The QE is set up with a control group and the following test groups:

- demote comments with v1 vaccine hesitancy classifier score of 0.75 (p59) or above with demotion strength of 0.01
- demote comments with v1 vaccine hesitancy classifier score of 0.75 (p59) or above with demotion strength of 0.001
- demote comments with v1 vaccine hesitancy classifier score of 0.8 (p68) or above with demotion strength of 0.01
- demote comments with v1 vaccine hesitancy classifier score of 0.8 (p68) or above with demotion strength of 0.001

Comments are only being demoted if they:

- (a) appear under posts marked with the vaccine subtopic by FIT,
- (b) have a v1 vaccine hesitancy classifier score of threshold or above,
- (c) are written in english and
- (d) are longer than 5 characters.

This is the evaluation process for which we got the thresholds for p59 (r17) and p68(r10). According to scuba logs, at least 5% of comments (100k+ per day) posted under vaccine-related Facebook posts express vaccine hesitancy and will be eligible for the demotion.

HYPOTHESIS

We hypothesize that downranking these comments will result in a decrease in engagement (specifically vpv and reactions) as well as USIs (hide_report_not-self delete). As a stretch goal we will monitor any increase of engagement on non vaccine hesitant comments.

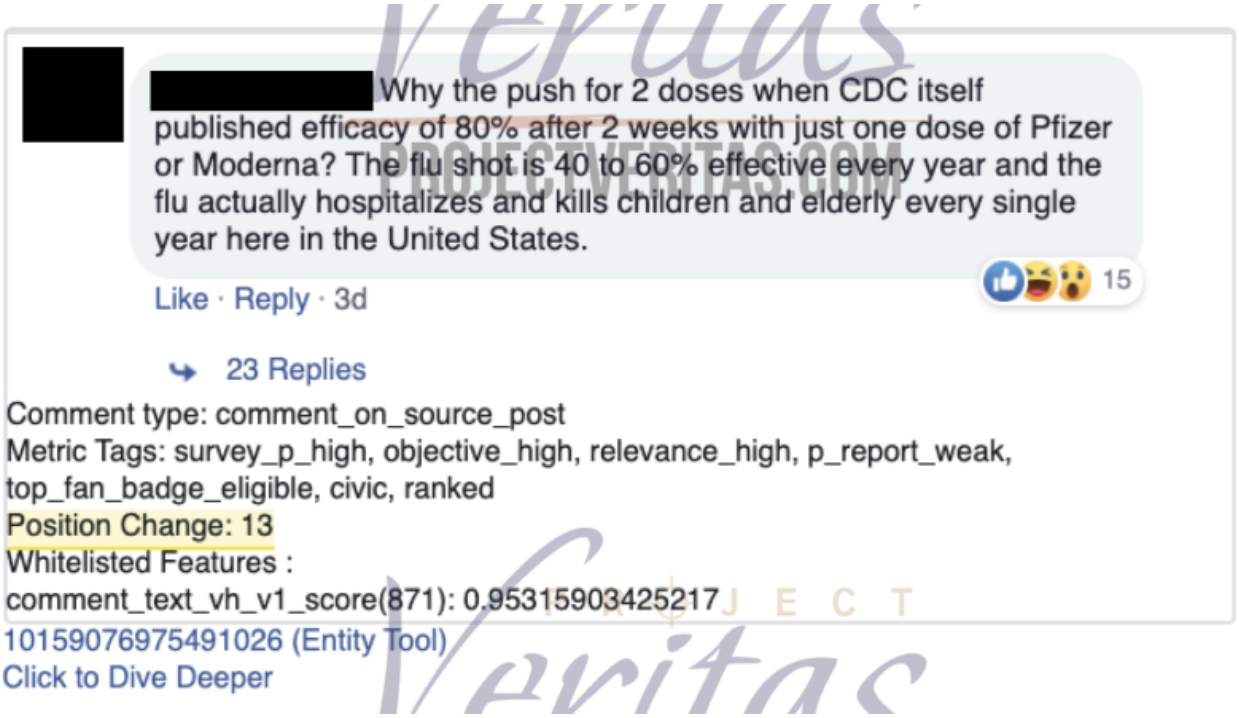
Cc Matt Gilles, Nick Gibian, Amit Bahl, Matt Motyl

— with Joo Ho Yeo.

6. One part of the document admits to "demoting" comments and reducing the rate of likes while conducting a test of 1.5 percent of the company's audience.

2. Change in "likes"
   a. For the vaccine hesitant comments we are demoting, we are reducing -2.64K likes (-42.5%), which would be 176K likes since the test size is 1.5%
      i. https://fburl.com/scuba/ranked_comment_actions/oldlnwjv

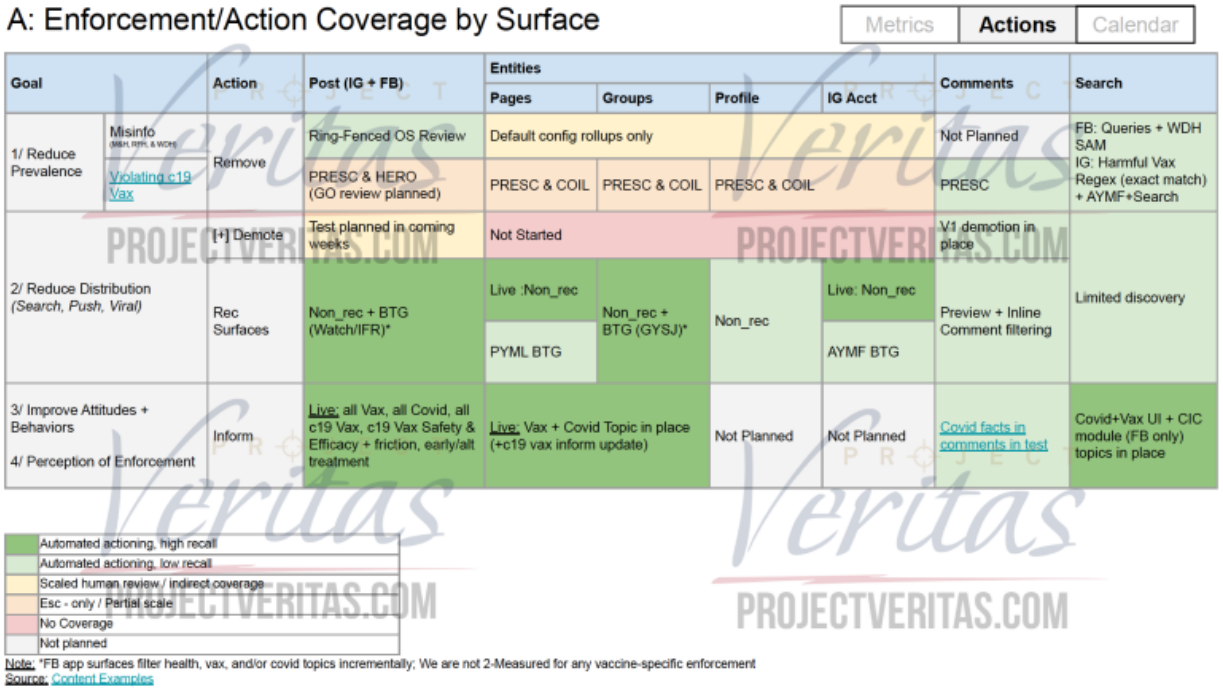
7. Other posts discussing the unique double dose of certain vaccines were described as "objective_high" and outlined a "position change."



8. Descriptions of the type of content that falls into Tier 0, Tier 1 and Tier 2.

Tier	Tier Name	Description	Mapped to B2V Tiers
T0	Violations of Policies: Coordinating Harm	Vaccine Interference: Coordinating (statements of intent, calls to action, representing, supporting or advocacy) OR depicting, admitting to, or promoting interference with the administration of a vaccine, including an event, group, page, account, etc dedicated to this purpose. Vaccine Explicit Discouragement: Calls to action, advocating, or promoting that others not get a vaccine, including an event, group, page, account, etc dedicated to this purpose. It is only when calling for *unspecified groups of people* to refuse a vaccine; not specific individuals or groups of people (i.e. "you", "Sarah", or "the elderly")	
		B2V Tier 1	B2V Policy
T1	Alarmism & Criticism	Criticizing Choice to Receive/Provide Vaccines: Disparaging others on the basis of their choice to vaccinate, or on their choice to vaccinate others	B. Criticizing Choice to Receive/Provide Vaccines
		Exaggerated Conclusions/Denialism: Content about vaccines and vaccination that suggests or implies that vaccines are unsafe, ineffective, sacrilegious or irrelevant Conspiracy Narratives: Content using conspiratorial language in reference to vaccination efforts where it suggests there is some purposely hidden widespread health harm, secret, or truth that people are being let in on	A. Sensational or Alarmist Vaccine Content
		B2V Tier 2	B2V Policy
T2	Indirect Discouragement	"Shocking, possibly true" Unproven or Severe Side Effects or Death: Content that discourages vaccination based on personal anecdote OR news articles of unproven or severe vaccine side effects, including claims of death	D. Shocking Stories
		Promotion of Vaccine Alternatives or Rejection: Content that directly or indirectly discourages vaccination through promotion of vaccine alternatives or celebration of those who refuse vaccination	C. Promoting Vaccine Refusals & Alternatives

9. An enforcement document obtained by one of the Insiders showed Facebook's goal of reducing the distribution of comments seemingly skeptical of Covid vaccination.



A: Enforcement/Action Coverage by Surface

Metrics | **Actions** | Calendar

Goal		Action	Post (IG + FB)	Entities				Comments	Search
				Pages	Groups	Profile	IG Acct		
1/ Reduce Prevalence	Misinfo (MSH, RPH, & WDH)	Remove	Ring-Fenced OS Review	Default config rollups only				Not Planned	FB: Queries + WDH SAM IG: Harmful Vax Regex (exact match) + AYMF+Search
	Violating c19 Vax		PRESC & HERO (GO review planned)	PRESC & COIL	PRESC & COIL	PRESC & COIL		PRESC	
2/ Reduce Distribution *(Search, Push, Viral)*		[+] Demote	Test planned in coming weeks	Not Started				V1 demotion in place	Limited discovery
		Rec Surfaces	Non_rec + BTG (Watch/IFR)*	Live :Non_rec	Non_rec + BTG (GYSJ)*	Non_rec	Live: Non_rec	Preview + Inline Comment filtering	
				PYML BTG			AYMF BTG		
3/ Improve Attitudes + Behaviors 4/ Perception of Enforcement		Inform	Live: all Vax, all Covid, all c19 Vax, c19 Vax Safety & Efficacy + friction, early/alt treatment	Live: Vax + Covid Topic in place (+c19 vax inform update)		Not Planned	Not Planned	Covid facts in comments in test	Covid+Vax UI + CIC module (FB only) topics in place

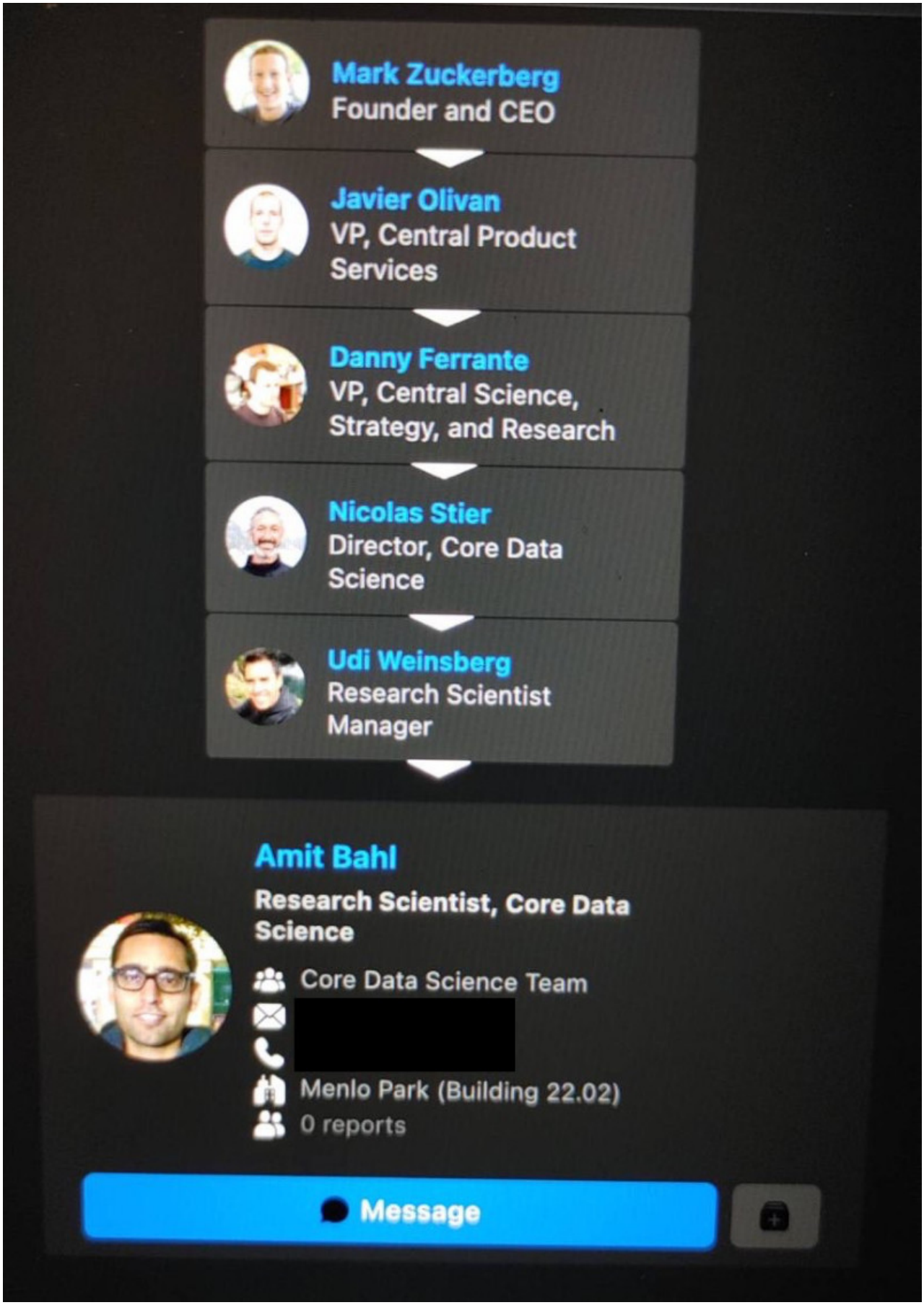
	Automated actioning, high recall
	Automated actioning, low recall
	Scaled human review / indirect coverage
	Esc - only / Partial scale
	No Coverage
	Not planned

Note: *FB app surfaces filter health, vax, and/or covid topics incrementally; We are not 2-Measured for any vaccine-specific enforcement
Source: Content Examples

10. Facebook document details how many languages their "Vax Safety" policy will be able to detect for suppression.

Note: All informs Covers FB + IG, only FB versions shown
Source: Inform Tracker; Inform Treatments+Reshare Friction Monitoring

11. Facebook's "Vaccine Hesitancy" project leaders listed in company's chain of command, reporting directly to Mark Zuckerberg.



12. Facebook's Ben Freeman sends internal message to colleagues informing them of data gathering and action to be taken regarding COVID-19 vaccines.

Health Integrity FYI
Open Group · 777 members · Major announcements and initiatives impacting 2+ teams outsi..

Posts Files Topics More

Ben Freeman
April 30 at 9:13 AM

[LAUNCH] CORRECT THE RECORD FOR INTERACTORS OF COVID-19 M&H ON IG

TL;DR:

- We launched Correct The Record (CTR) notifications to 5% of IG App users in all markets on 29th April and plan to launch it to 20% by 7th May.
- Users who interacted with COVID-19 content we later removed for M&H policy violations will receive an activity notification (seen in app when you tap the heart icon) with contextual information to "correct the record" (see below for screenshots). For extreme clarity; this is not a push notification.

BACKGROUND

- In May 2020, we launched in-feed modules (QP) on FB and IG targeting users who commented on/liked/shared harmful misinformation about COVID-19 with content from WHO busting common COVID-19 Myths. In Oct 2020 we launched CTR 2.0 on FB (Jan 2021 on Msite and FBLite) that triggers a notification to the user upon deletion of content for COVID-19 M&H. This launch is of notifications to IG users.
- Since launch of CTR v2.0, we have reached ~9m users on FB with a click through rate of 35.5%.

LAUNCH DETAILS AND TIMELINES

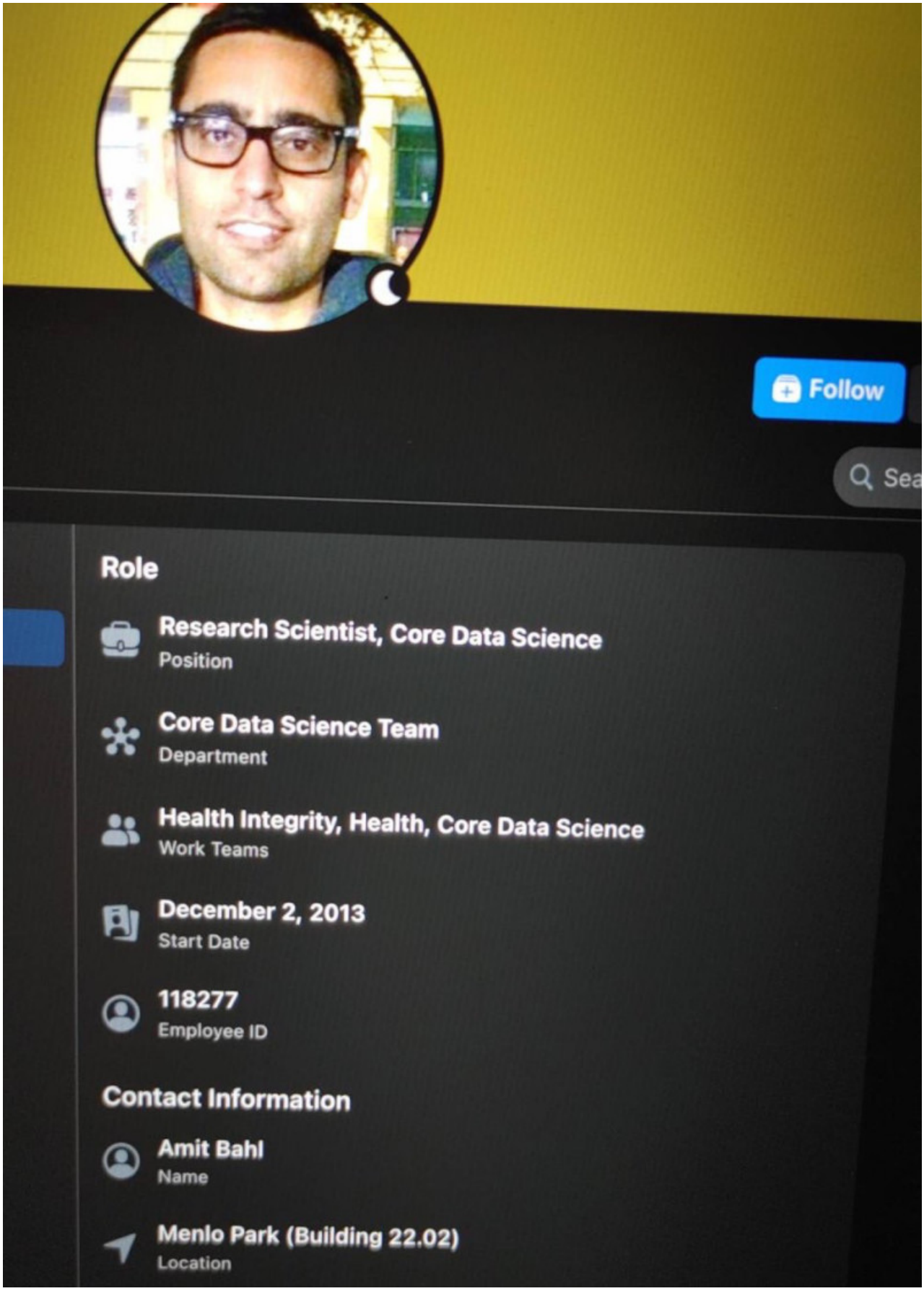
- We launched to 5% IGApp user globally on 29th April
- At 100% we expect to be sending out 100k notifications week. We estimate that only 1% of users we reach out to will receive more than one notification each day.

METRICS STRATEGY:

- Success metrics:
  - CTR to open notification > 5%
  - CTR to suggested actions (of all impressions of the notif) > 5% (QP baseline)
- Counter-metrics:
  - Notification delete/dismiss rate
- Survey metrics:
  - We are not planning on running a dedicated survey for this feature at the moment. Limitations in the IG survey tool do not allow for inclusion of specific posts and questions tailored to the included post. We do not think a survey with these limitations will give us actionable learnings.

COMMS STRATEGY:

13. Amit Bahl's job role within Facebook explained.



14. "Vaccine Hesitancy Comment Demotion" Tier 1 and Tier 2 explained.

Tiering Summary:

Tier 1: Alarmism & Criticism

A. Sensational or Alarmist Vaccine Content: Suggesting that vaccines are unsafe, ineffective, sacrilegious or irrelevant, in exaggerated, conspiratorial, or sensational terms
B. Criticizing Choice to Receive/Provide Vaccines: Disparaging others on the basis of their choice to vaccinate, or on their choice to vaccinate others

Enforcement Principle: This content could present a barrier to vaccination in many contexts

Tier 2: Indirect Vaccine Discouragement

C. Promoting Vaccine Refusals & Alternatives: Implicitly discouraging vaccination by advocating for alternatives or celebrating those who refuse vaccination
D. Shocking Stories: Potentially or actually true events or facts that raise safety concerns, indicated by sharing personal anecdotes or news events of severe adverse events in hyperbolic terms or without context

Enforcement Principle: This content could present a barrier to vaccination in certain contexts, particularly in entities sharing high rates of it.

15. Facebook describes how to label different types of posts.

Labeling Guidelines

Look at the post and spend 30-60 seconds to make sure that you understand it. It's important to consider any text in conjunction with any media - for example, you'd want to think about how a photo might be interpreted *in combination with* any text that's present. In addition to the text itself, if the posts contains...

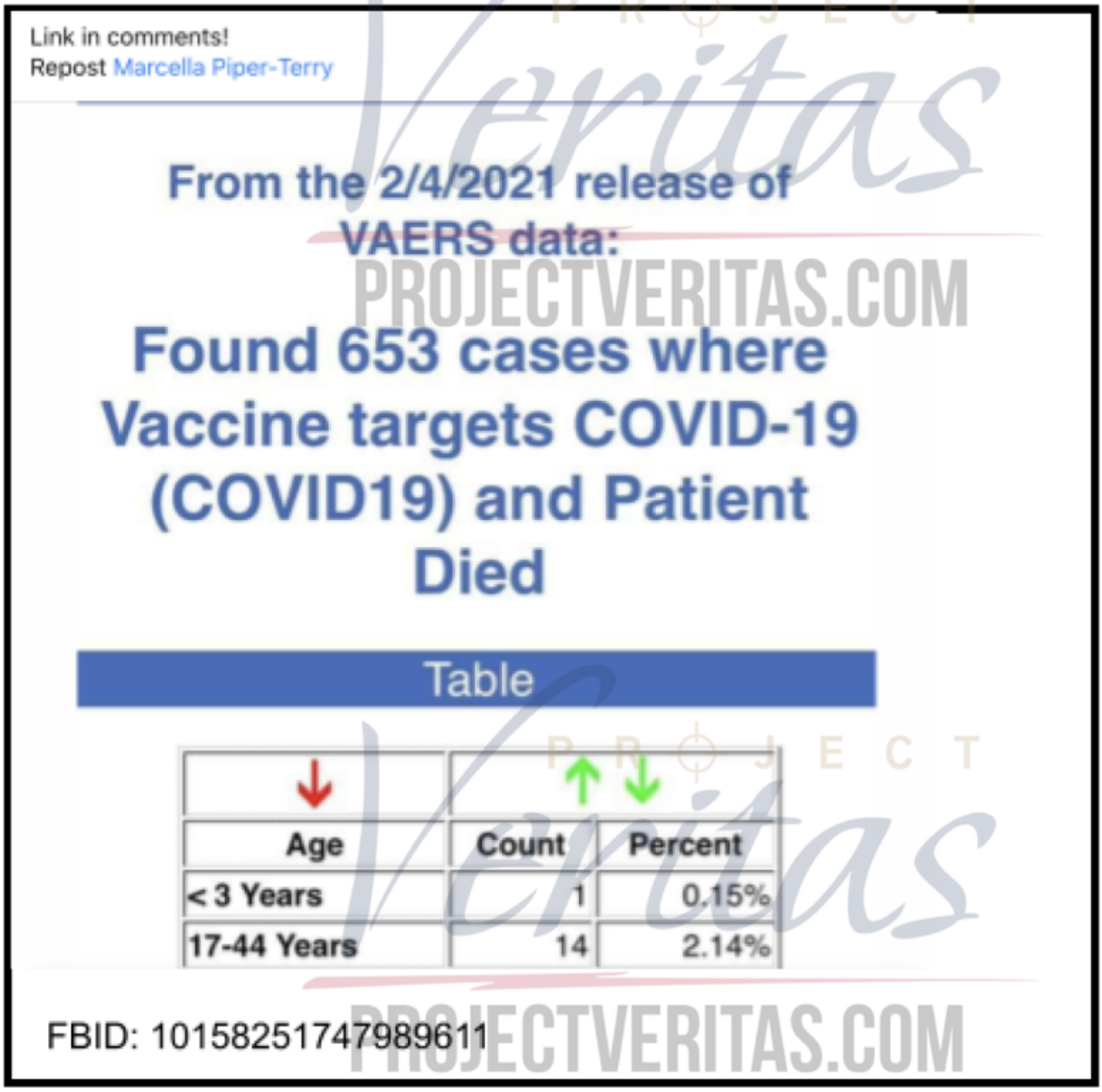
1. **...a photo**, consider the photo(s) (including text overlaid on it), as well as any text appearing above the photo.
2. **...text only**, consider the text (including emojis) in the post.
3. **...a video**, watch the video for up to 60 seconds (if the video is long watch first 10-15 seconds, click through 5 spots and watch 5-10 seconds at each) then also consider any accompanying text.

16. Facebook elaborates on Tier 2's "Indirect Discouragement."

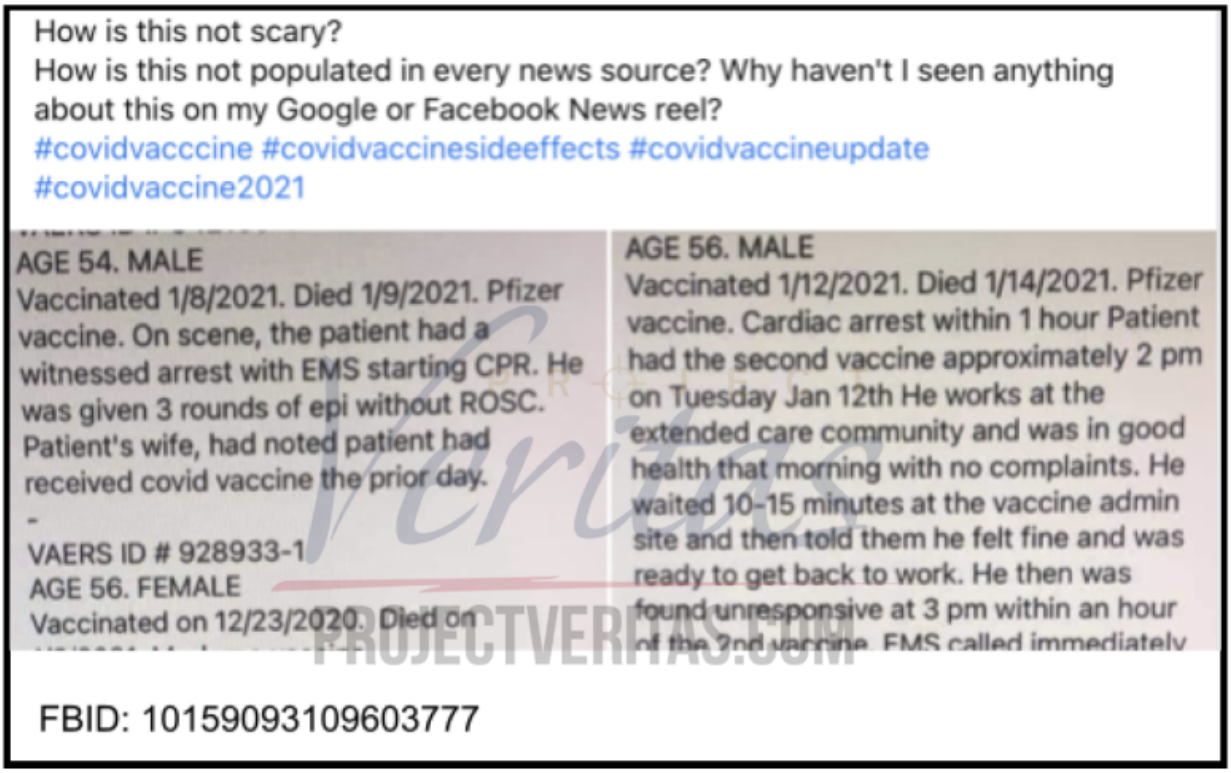
[Tier 2] Indirect Discouragement

- Promotion of Vaccine Alternatives: Promotion of vaccine alternatives
  - INCLUDES
    - Promoting alternatives to vaccination, such as:
      - ☐ Suggesting that getting COVID-19/natural immunity is 'better' for another person vs. getting the vaccine
        - "*Just skip the vaccine and trust in herd immunity*"
      - ☐ Minimizing the risks of the disease against which you can get vaccinated
        - Suggesting that vaccines aren't necessary given low COVID death rates

17. VAERS data analyzed by Facebook.



18. Vaccine Hesitancy comments targeted by Facebook.



19. Facebook describes what it classifies as "Shocking Stories" in regard to vaccines.

D. Shocking Stories: Potentially or actually true events or facts that can raise safety concerns, indicated by sharing the following either without context or with hyperbolic terms:
   - Content pointing to VAERS (Vaccine Adverse Events Reporting System) data that suggest extreme risks without providing full context
   - Content that pairs a shocking story with vague statements such as "look at what happened after she got vaccinated"
   - Content describing individual/personal stories of extreme safety risks or health harms after vaccination that may undermine or discourage vaccination
     - *"Excruciating pain after my second vaccine! Shaking so bad almost to convulsions"*
   - Content citing news articles about individuals experiencing adverse events after receiving a vaccine
     - *"an individual in california who tested positive for covid-19 in late december and was vaccinated thursday died several hours after receiving the shot"*
     - *"Deaths have reached 653 as of February 14, 2021. All of these people took the vaccine at their own risk. :'("*
   - *Content referencing expected mild side effects would be labeled as Neutral (not shocking), such as:*
     - *Redness, swelling, and/or soreness on the arm where you received your shot*
     - *Fever, Chills, Headache, Fatigue/tiredness throughout the rest of your body*

Effect on Drivers of Intent to Vaccinate: (A) Intent: Emotions. Key Beliefs, Social Norms

Read the entire Facebook Vaccine Hesitancy Comment Demotion document here.

Read the entire Facebook Global Operations Primer - Health Misinformation document here.

BE BRAVE. Do Something. Stand up. The time is now. If you witness corruption, fraud, waste, or lying within Big Tech, private entities, or the government — things the public needs to know — Send an encrypted message to our Signal: 914-653-3110. Or contact us at VeritasTips@protonmail.com

About Project Veritas

James O'Keefe established Project Veritas in 2011 as a non-profit journalism enterprise to continue his undercover reporting work. Today, Project Veritas investigates and exposes corruption, dishonesty, self-dealing, waste, fraud, and other misconduct in both public and private institutions to achieve a more ethical and transparent society. O'Keefe serves as the CEO and Chairman of the Board so that he can continue to lead and teach his fellow journalists, as well as protect and nurture the Project Veritas culture.

Project Veritas is a registered 501(c)3 organization. Project Veritas does not advocate specific resolutions to the issues raised through its investigations. Donate now to support our mission.

HELP EXPOSE CORRUPTION

SEND A TIP

DONATE NOW

1214 Boston Post Road No. 148
Mamaroneck, NY 10543

(914) 908-2300