Last updated: Apr 21, 2021
Status: Draft
Questions: [redacted]@

Addressing viral suicide & self-injury challenges across the FOA

As we see more SSI challenges emerging and continued media interest in them, we are looking to

- (a) share facts around the nature of viral challenges with established SSI partners, to empower them to create resources educating their communities, and
- (b) develop standalone resources on the FB Safety Center with guidance for young people, parents, and educators on what to do if confronted with a viral challenge.

Trends around viral challenges

An SSI viral challenge is an event that promotes or encourages others to complete a challenge related to SSI, including self harming or suicide.

How often do we see viral SSI challenges on Instagram?

We have noticed an increasing rate of SSI viral challenges on the platform.

- 2016 - 2020 H1: 2 challenges
  - Blue Whale: A 'group administrator' assigns 'daily tasks' to members, which they have to complete for 50 days e.g. self-harming, watching horror movies and waking up at unusual hours. These gradually get more extreme; on the 50th day, the controlling manipulators behind the game reportedly instruct the members to commit suicide
  - Momo: You have to do whatever "Momo" — or the person sending you the Momo image — says, or else Momo will come for you and your family and curse you. The final step in the process is killing yourself and filming it
- 2020 H2: 3 challenges
  - Dog Face ([HYPERLINK "https://www[redacted]" \h]): Accounts encouraged others to engage in "challenges," which solicited photo proof of inflicted self harm, otherwise threatening to reveal users' private information, or threatening to kill users and/or their family members
  - 6inner ([HYPERLINK "https://www.[redacted] \h], [HYPERLINK "https://www[redacted]" \h]): Variant on Blue Whale Challenge; a stranger messages a person on Instagram, giving them "challenges" or "games" to participate in, collecting personal & sensitive information from them, and ultimately threatening with releasing the sensitive information if they don't comply with the challenges

Redacted - CSI

- 2021 H1 to date:
  - White Rat ([HYPERLINK "https://www[redacted]" \h]): An iteration of the Blue Whale Challenge

- March 3 ([HYPERLINK "https://www[redacted]" \h]): A trend/viral challenge on TikTok encouraging teens to commit suicide on 3 March, that we were concerned might spill over to Instagram
- Apostle of God ([HYPERLINK "https://www[redacted]" \h]): An anonymous figure on Instagram making people draw a "pact of fidelity", followed by numerous tests of loyalty, increasingly pressing and dangerous requests. A 15 year old girl in Italy nearly threw herself off the roof of a building to acquiesce to the Apostle's demands
- Resurgence of Dog Face ([HYPERLINK "https://www[redacted]])

How do we come to learn of them?

- We usually come to learn of viral challenges from news / media outlets reporting on them, which FB employees, Risk & Response, Press Escalations, or Markets teams then flag to our teams
- Alternatively, FB employees will share reports of a challenge from family/friends. Examples:
  - [HYPERLINK "https:/[redacted]]
  - [HYPERLINK "https:/[redacted]]

What are the general outcomes of proactive sweeps (do they yield a lot of actionable content or?)

- It depends in part on how viral/prevalent the SSI challenge/event is and the nature of the challenge (i.e. do we deem a certain type of photo as actionable content, like we did with Dog Face, or do we need a combination of photo and additional signal to action it?).
- For example, Dog Face's proactive sweep resulted in thousands of content match hits and we ended up bulk actioning them through the sweep. On the other hand, the White Rat sweep resulted in less than 100 pieces of actionable content pieces because there was no photo involved and we were going strictly off of keyword matches.
- Publicly, the profiles usually present as representative of the challenge
  - e.g. it's not "Matt Denton tells you to try the Momo challenge", it's an account called "Momo challenge"
- The behaviour usually starts on a public surface then migrates to a private surface - people push the challenges in private conversations

Do we find much content that may lead to imminent harm?

- We have not encountered much content that posed significant imminent harm; while we have found content that violates our policies, our policies around suicide challenges also lean towards the most cautious approach to challenges and content removal.
- A tiny % of challenge-presenting profiles are actually engaging in the challenge (based on our reviews of inbox conversations) - a huge number of them engage in copycat behaviour

Last updated: Apr 21, 2021
Status: Draft
Questions: [redacted]@

- The majority of accounts we action for viral challenges aren't actually engaging in any abusive behaviour besides representing as being a challenge
  - e.g. Account named as "dogface" but not actually doing anything violating
- Our policies consider these accounts to be inherently violating but the harm they actually result in is debatable. Many of these accounts are children engaging in pranks e.g. you hear about this challenge, you decide to create a profile of your own, you reach out to a classmate you don't like

- We take down this content because of external perceptions of user safety rather than actual observed harm happening on the platform. Of all the suicide harms that happen on the platform, this is the one that proliferates because of awareness of the harm on our platforms

What have we done about viral SSI challenges?

- [HYPERLINK "https:/[redacted]]: We have developed a standard framework for managing SSI escalations and viral trends. It covers best practices that SSI Process has developed based on lessons the team has learned from a number of SSI trends that have surfaced
- [HYPERLINK "https://www[redacted]n] - Launching Known Question on March 31 to define and clarify how we think about SSI Viral Events on our platforms
- We've historically taken a reactive approach to dealing with viral events (i.e. we deal with it once it's become so viral that media outlets catch wind of it). It's very difficult to predict/surface trends before they become viral, so it's tough to proactively ID them.
- In an effort to move the needle towards proactive detection of trends (before they become viral), we shipped new guidelines (see below) with SSI Viral Events signals in early March. We're hoping that this will lead to reps flagging early signs of viral events, but we have yet to see the results of it since we haven't yet had enough time to see any insights come through yet

Any other observations?

- Increasingly we get enquiries about and see evidence of cross-app contagion, particularly incoming challenges that originated on TikTok
- Proactive Detection using classifiers is also usually not an option. The time it takes to build a classifier is typically too long to respond to these challenges before they spread. Additionally, the limitations on SSI on creating profile level detection in the EU would also block us.

Commented [1]: some additional observations - targeting minors?

Commented [2]: This intuitively seems right but @[redacted]@fb.com would you say this is the case?

Commented [3]: I would say this is a fair assumption but we haven't done analyses specifically around the age of users who take part in these challenges @[redacted]@fb.com

Efforts to Proactively Identify SSI Viral Events on the Rise

1. If you receive 3+ jobs within a day that each contain a common "suspicious indicator", continue to Step 2. We define suspicious indicators as content that contains one or more of the following in the format of photos, usernames, bios, or captions:
   a. A distinct symbol or character (for example, a dog face image, a whale emoji, etc)
   b. Variations of a distinct name (e.g. Jonathan Galindo, Blue Whale, Momo, 6inner)
   c. Username that contains "666"
   d. References to a "game", "play", "text", or "DM", *together with additional signals*. Refer to "List of Signals" section below for additional details

   i. Suspicious content often but not always undertakes the form of an IGPR with a Bio containing sentences such as "wanna play a game", "Let's play a game"
2. Flag to your site's POC for trend flagging via Vendor Portal or SRT (based on your site/vendor's reporting protocol) with the following information:
   a. Job ID(s): It is expected that you may not have taken note of the previous 2+ jobs you reviewed. Even if you have only one job ID available, please report it by listing the available job ID(s) and adding a description of the number of similar jobs you encountered.
   b. The suspicious indicators or Signals associated with these jobs
   c. An approximation of how many times you have encountered a similar job

List of Signals

Signals that could be found in Username and/or Bio				
Signal	**Example 1**	**Example 2**	**Example 3**	**Example 4**
Emojis	the devil emoji 😈	the cross emoji ❌	the knife emoji 🔪	the blue whale emoji 🐋
Username and Bio Keyword 1: "666"	Do not talk to me 666 If you dare there is no turning back 666	666 recomende 666	666 no te recomiendo que me hables 666\n Hip Op Hi (translated: 666 I don't recommend that you talk to me 666'	Usernames such as _jonathan.galindo_666_
Username and Bio Keyword 2: "game"	Play games	If you don't want to play my game, I will find you.	let's play a game	I'M JONATHAN AND I'M LOOKING FOR FRIENDS :)\nWRITE ME AT 00:00 TO PLAY A GAME
Bio Keyword 3: "No me hables" / "Don't talk to me"	No me hables❌Dont call me❌	No me hables o te irá peor 666 (translated: Do not talk to me or it will be worse 666)	No me hables o te mato (translated: Don't talk to me or I'll kill you)	
Bio Keyword 4: "wanna play a game"	Dm ne if you wanna play a game	do you wanna play a game?	wanna play a game?	
Bio Keyword 5: "DM" + additional signal (e.g. SSI emojis, 'challenge', 'game')	my dms are open	dm if you want to play	Dm me to play a very fun game	
Bio Keyword 6: "text" + additional signal (e.g. SSI emojis, 'challenge', 'game')	text me at 3 am and you will be answered	Send me a text and you will see	If u want to play , text me	
Username and Bio Keyword 7: "desafio"/ "challenge"				
Username and Bio Keyword 8: bluewhale	bluewhale.jonathangalindo	bluewhalejonathangalindo	bluewhale_challenge_f57	