Online Communities of Adolescents and Young Adults Celebrating, Glorifying, and Encouraging Self-Harm and Suicide are Growing Rapidly on Twitter

PRESENTED BY

RUTGERS
Miller Center for Community Protection and Resilience

RUTGERS
Center for Critical Intelligence Studies

NC LABS at RUTGERS

POWERED BY

NCRI
NETWORK CONTAGION RESEARCH INSTITUTE

Alex Goldenberg, Corresponding Author
alex@ncri.io
Lead Intelligence Analyst, Network Contagion Research Institute; Research Fellow, Miller Center for Community Protection and Resilience, Rutgers University

John Farmer, Author
Former New Jersey State Attorney General and Senior Counsel, 9/11 Commission; Director, Miller Center for Community Protection and Resilience, Rutgers University

Lee Jussim, Ph.D., Author
Former Chair, Distinguished Professor, Department of Psychology, Rutgers University

Loree Sutton, MD, Author
Psychiatrist; retired US Army Brigadier General; Strategic Advisor, NCRI

Danit Finkelstein, Author
Graduate Student, Department of Psychology, Rutgers University

Cristian Ramos, Author
Analyst, Network Contagion Research Institute

Pamela Paresky, Ph.D., Author
Visiting Fellow, Johns Hopkins University SNF Agora Institute; Senior Scholar, Network Contagion Research Institute

Joel Finkelstein, Ph.D., Author
Chief Science Officer and Director, Network Contagion Research Institute; Senior Research Fellow, Miller Center for Community Protection and Resilience, Rutgers University

NCRI Insights Report

Online communities of adolescents and young adults celebrating, glorifying, and encouraging self-harm and suicide are growing rapidly on Twitter.

Key Takeaways

- A community promoting self-harm (specifically, “cutting”) is circulating graphic and bloody depictions of self-injury on Twitter.
- In October, 5Rights, a UK-based children’s digital rights charity, alerted Twitter that their algorithms were pushing those searching the terms “self-harm” to profiles promoting self-harm rather than profiles connected to finding help.
- Since October, the use of hashtags related to self-harm (e.g. like “#shtwt” for **S**elf-**H**arm **TW**i**T**ter) has increased roughly 500%, averaging tens of thousands of mentions per month. Many community members appear to be adolescents and young adults.
- *Hashtags associated with ``shtwt” are peaking as well. These terms are usually associated with and accompanied by photos of severe and even potentially life-threatening self-inflicted wounds. These images and the cutting behavior are praised, celebrated, and encouraged.*
- *The vast majority of this content is in direct violation of Twitter’s Suicide and Self-harm policy, which states “users may not promote or encourage suicide or self-harm,” and its Sensitive Media policy which prohibits depictions of “gratuitous gore, bodily fluids such as blood, and serious physical harm, including physical wounds."*
- Evidence suggests adult online predators are likely engaging with these communities.
- These data appear to be the tip of the iceberg. The NCRI has identified a number of rapidly growing, and in some cases overlapping, Twitter communities dedicated to the glorification of eating disorders, mass shootings, and more.

Executive Summary

Twitter hosts a massive community that glorifies and encourages self-harm — specifically "cutting." Graphic photographs of what appear to be bloody self-injury by people who have sliced into their skin continues to proliferate, many such tweets garnering unusually high engagement given the small number of followers of the posting account. Photographs and other images are accompanied by slang terms for blood as well as for the depth, pattern, and complexity of cuts. Photographs depicting wounds that are bloodier and more severe, more dangerously deep, and more complex in number and/or design of cuts are more widely circulated than those that depict less serious wounds.

In October 2021, Twitter was alerted to the presence of the hashtag "#shtwt" (short for **S**elf-**H**arm **TW**i**T**ter), and that their algorithms were pushing those searching the terms "self-harm" to profiles in this community promoting self-harm rather than profiles connected to finding help. Since then, despite Twitter having claimed it would take action against tweets that violate their rules on suicide and self-harm, the use of related hashtags has seen exponential growth, and mentions of shtwt are up 500%.

The vast majority of the content circulating within the shtwt community is in direct violation of several Twitter policies. The NCRI has also identified a number of dangerous, and in some cases, overlapping communities that encourage and glorify eating disorders (in particular, a fixation on shedding weight until the skeleton is visible), mass shootings, and more.

Background

Prior to the birth of online internet forums in the 1990's an individual engaging in non-suicidal self-harm (NSSI) had few opportunities to discuss her or his habits other than with friends, family, and medical or mental health professionals. Today, platforms such as Tumblr, Reddit, and Twitter are home to hundreds of communities of adolescents and young adults engaging in NSSI who "turn to the internet to find information and receive validation and social support" according to a qualitative analysis of social media and self-harm photographs in Frontiers Psychiatry.[1]

Some of these communities are tightly moderated and operate primarily as vehicles for people engaging in self-harm to get help. For example, some serve as a resource to communicate and relieve distress, acting as a support network that encourages people engaging in harmful behaviors to get help.[2] Online communities serving such functions are not the focus of this report. Our focus is Twitter communities that celebrate, glorify, validate, and encourage those who engage in self-harm to continue and escalate self-injurious behavior.

[1] https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00274/full
[2] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181722

The acute dangers of social media use for minors have received a great deal of attention in the psychiatric and psychological academic literature. Social media use by minors across most common platforms (Facebook, Twitter, Instagram, etc.) has been steadily increasing over the past decade.[3] Some studies find that social media use is associated with depression and suicidality,[4] There has also been a significant increase in nonsuicidal self-injury (NSSI) a socially contagious[5] phenomenon that is too common[6] among adolescents, and more prevalent among girls than boys. For girls[7], NSSI is more likely to involve cutting than it is for boys.

SHTWT Communities: Celebrating and Gamifying Self-Harm

Online communities that celebrate psychopathology and self-harm are not new. In these cases, the online forum acts as an incubator; an environment that both normalizes the behavior and creates a culture that is both supportive and competitive. It is supportive in that it affirms and celebrates cutting. It is competitive in that it encourages people in distress—including adolescents—to increase the severity of their self-inflicted wounds as part of a process of "gamifying" self-harm.

The NCRI has charted the growth of a number of highly active subcultural communities using coded language related to self-harm since the beginning of the pandemic. The self-harm community is identified by specific hashtags and memes. The term "shtwt" (with or without the #) is commonly used to communicate "self-harm Twitter." Coded language includes labels for layers of skin, types of injuries, and patterns of resulting blood. These are often benign or even cute words such as "catscratch" for the superficial self-cutting that often looks like cat scratches; "beans" (which refers to cutting deep enough that one gets to the subcutaneous layer, which, as the image reveals, has the appearance of beans); "armgills" (as if cuts are like gills on fish); and "raspberry filling" to refer to blood. Related coded terms include "moots" (for "mutuals," as in "mutually engaging in self-harm"); "ed" (for eating disorders) — sometimes with "moots" or "twt" (for "Twitter"); "ouchietwt"; and blunter terms, such as "gore," "gorey," "gory," or "blood."

Such insider jargon fosters a sense of community in which encouragement to increase the depth and severity of the self-inflicted wounds (e.g. "go deeper") is perceived as positive and supportive rather than abusive and dangerous. Even a brief perusal of Tweets that tag

[3] Reid Chassiakos, Y. L., Radesky, J., Christakis, D., Moreno, M. A., Cross, C., Hill, D., ... & Swanson, W. S. (2016). Children and adolescents and digital media. *Pediatrics*, *138*(5).
[4] Memon AM, Sharma SG, Mohite SS, Jain S. The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. Indian J Psychiatry. 2018 Oct-Dec;60(4):384-392.
[5] Claes, Laurence, et al. "Brief report: The association between non-suicidal self-injury, self-concept and acquaintance with self-injurious peers in a sample of adolescents." *Journal of Adolescence* 33.5 (2010): 775-778.
[6] Monto, Martin A., Nick McRee, and Frank S. Deryck. "Nonsuicidal self-injury among a representative sample of US adolescents, 2015." *American journal of public health* 108.8 (2018): 1042-1048.
[7] Barrocas, Andrea L., et al. "Rates of nonsuicidal self-injury in youth: age, sex, and behavioral methods in a community sample." *Pediatrics* 130.1 (2012): 39-45.

themselves with these codes reveals photos illustrating how severely people have cut into themselves. Many also flag themselves as adolescents by posting their age and inviting peers their own age to interact with them on Twitter around self-harm.

The shtwt community revolves around celebrating self-harm and seeking validation. One highly engaged post from August 18th is captioned “this is the deepest I’ve done someone be proud of me,” and is accompanied by coded terms for the depth of the cuts and an image of the wounds. The post received over 2,000 likes, 165 retweets, and 80 comments. Users responded by saying “that’s so pretty,” “how beautiful” and, “what did you use.” Another highly engaged tweet, with over 2300 likes showcases an anime character daydreaming about self-harm. Since there is no way for an account to “like” a post more than once, each “like” represents a unique viewer. Two thousand likes is an unusually high level of engagement, even for an account with tens of thousands of followers. These accounts have less than three hundred followers.

Fig 1. A highly engaged tweet of self-inflicted wounds captioned “tonight’s first small section,” along with #shtwt and #raspberryfillingtwt

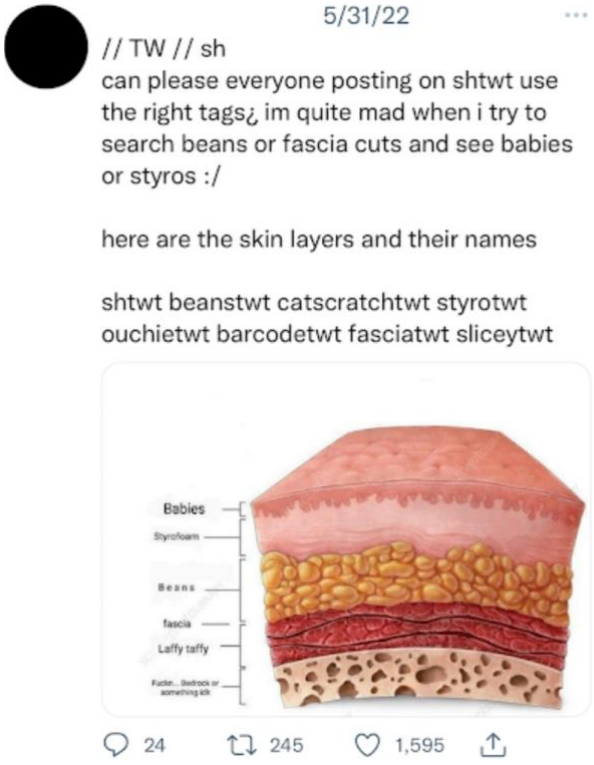


Fig 2. A tweet that garnered a high level of engagement included this image accompanied by “shtwt” term. This image illustrates different layers of skin labeled with their memetic codes. Hashtags and terms associated with layers of skin such as styrotwt, beanstwt, laffytaffytwt, and others are mentioned thousands of times per month allowing users to gamify and market self harm. Wounds associated with “beanstwt” tweets are extremely severe. These are often the tweets and images most liked and shared.

Fig 3. User claims to be heading to the hospital after claiming to “hit beans in one swipe.”

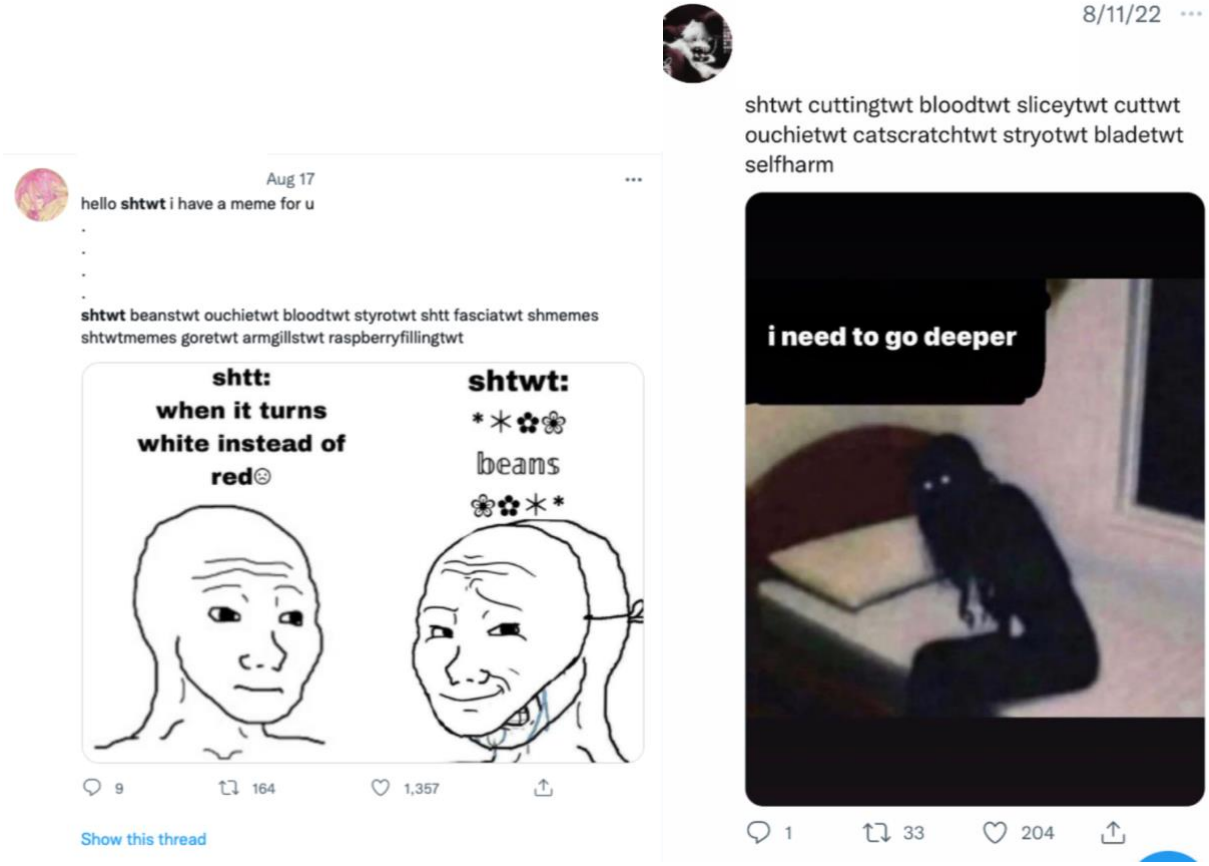


Fig 4. Example of highly engaged tweets associated with the hashtags. Image on the left represents the translation of code for deep cuts that reveal the subcutaneous layer of fat from another online platform (“shtt” is “self-harm tiktok”) to Twitter. Image on the right shows a meme with the caption “I need to go deeper.”

Concerning Community Growth Despite Formal Warnings

In October 2021, 5Rights, a UK-based children's digital rights charity, submitted the findings of their investigation to the Information Commissioner's Office. Dozens of tech companies, they concluded, are "systematically endangering children online" by facilitating children's ability to connect with others celebrating self-harm. Their research identified a specific problem with Twitter's algorithmic recommendation system: *It was steering accounts with child-aged avatars searching the words "self-harm" to Twitter users who were sharing photographs and videos of cutting themselves.* 5Rights also highlighted content associated with #shtwt, #ouchietwt, and #sliceytweet, which was in direct violation of Twitter's content policy.

In response to the report, Twitter told the *Financial Times* that it had blocked #shtwt, #ouchietwt, and #sliceytweet from appearing in "any future trends on the app." But the damage had already been done. Twitter reiterated that "it is against the Twitter rules to promote, glorify or encourage suicide and self-harm" and claimed, "Our number-one priority is the safety of the people who use our service. If tweets are in violation of our rules on suicide and self-harm and glorification of violence, we take decisive and appropriate enforcement action."

However, the NCRI has discovered that in the months since Twitter announced they were taking "decisive and appropriate enforcement action," the flagged community has seen prolific growth. The number of users with #shtwt in their bios has doubled since October 2021. Monthly mentions of "shtwt" and associated self-harm terms have increased by over 500% since Twitter was first publicly alerted to the issue. In October 2021 there were 3880 tweets (including retweets) using "shtwt." Over the past 6 months the average number of monthly tweets has been 20,000, reaching a peak of close to 30,000 in July 2022.

The use of related, more specific hashtags, such as beanstwt, a reference to extremely deep cutting—code for cutting into the subcutaneous layer of skin—continues to increase. In October, the number of tweets mentioning beanstwt was under 1000. In August, over 4500 tweets included the term. The community also has a number of new hashtags/terms and rudimentary methods to avoid detection and account suspension (including tweeting that the blood in accompanying photos is fake).

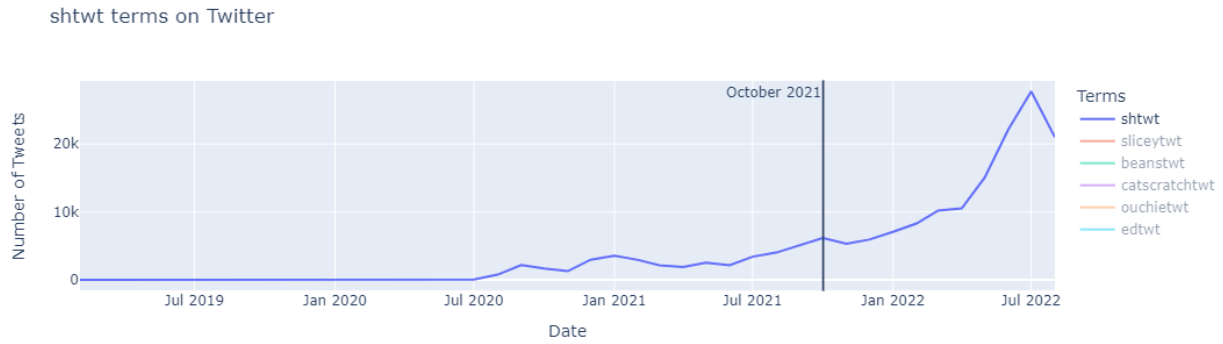


Fig 5. Timeline analysis of tweets and retweets mentioning the term “shtwt” from January 2019 to August 2022. The term started circulating at a low frequency in September 2019, and is now mentioned on average over 20k times per month.

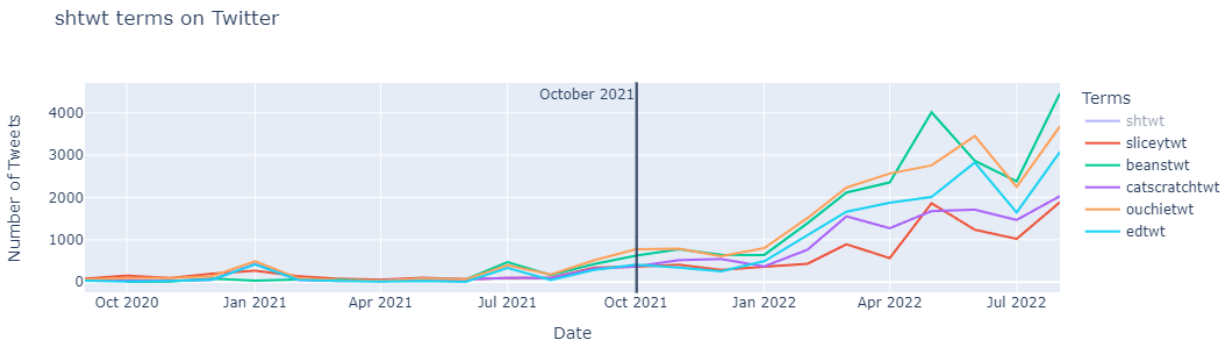


Fig 6. Timeline analysis of tweets and retweets mentioning self-harm associated terms such as “sliceytwt,” “beanstwt,” “catscratchtwt” and “ouchietwt” from October 2020 to August 2022. These terms are mentioned thousands of times per month.

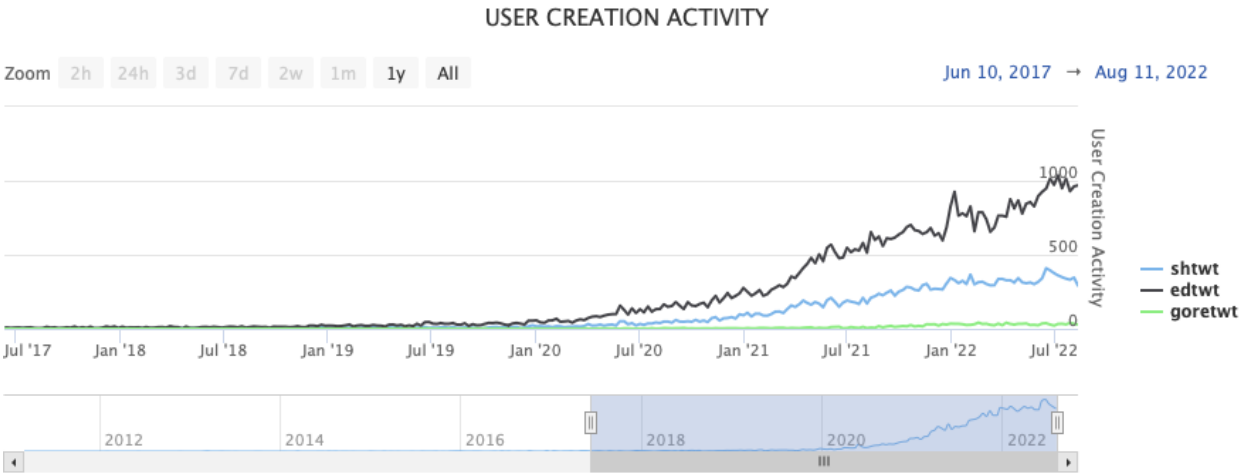


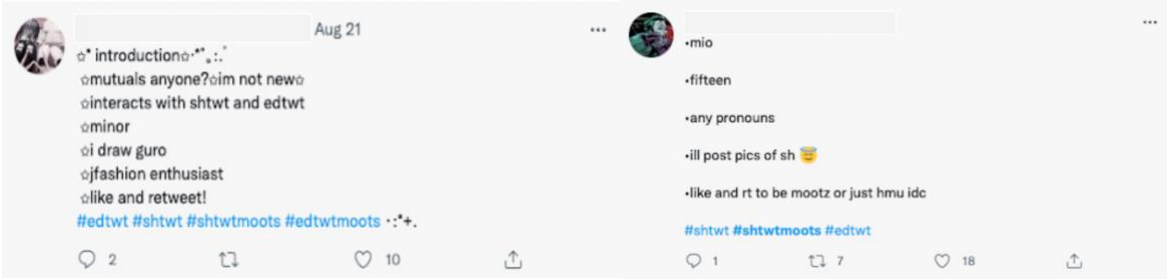
Fig 7. Timeline analysis of user-growth for users who include “shtwt,” “edtwt,” and “goretwt” in their Twitter bio descriptions.

Online Predators likely in Self-Harm Communities

There is a high likelihood that adult online predators are engaging in these communities. According to the FBI, there are roughly 500,000 online predators active on a daily basis, many of whom pretend to be teenagers. More than half of their victims are between the ages of 12 and 15.

The NCRI identified a number of accounts whose users claim to be minors, but whose profiles and tweets signal the operational security (OPSEC) sophistication of an internet-savvy adult. Some accounts are able to circumvent Twitter's algorithms OPSEC techniques like adding characters (for example “!!”) to obscure discord handles, adding spaces between letters (for

example h o r n y) to prevent automatic risk scoring or flagging, and including terms like “minor” in their tweets and/or bios to obscure their age. This is not amateur-level operational security and suggests that some of these accounts encouraging self-harm are run by trolls and predators.



Summary of current Twitter policy violation:

Twitter's official *Suicide and Self Harm*[8] and *Sensitive Media*[9] Policies clearly forbid the vast majority of the content circulating within the shtwt communities. The growth of self-harm communities results from Twitter's lack of enforcement of its own terms of service—rather than an oversight or omission in the terms of service themselves. (See appendix for examples.)

- ***Suicide and Self Harm policy violated by members of self-harm Twitter***
  - *"Under this policy, you can't promote, or otherwise encourage, suicide or self-harm.*
  - *Violations of this policy include, but are not limited to:*
    - *Encouraging someone to physically harm or kill themselves;*
    - *Asking others for encouragement to engage in self-harm or suicide*
    - *Sharing information, strategies, methods, or instructions that would assist people to engage in self-harm and suicide.*

- ***Sensitive Media policy violated by members of self-harm Twitter***
  - *"You may not promote or encourage suicide or self-harm."*
  - *"There are also some types of content that we don't allow at all, because they have the potential to normalize violence and cause distress to people who view them." Such sensitive media include:*
    - *"Graphic violence that depicts death, violence, medical procedures, or serious physical injury in graphic detail. Some examples include but are not limited to:*
      - *Physical child abuse*
      - *Bodily fluids including blood, feces, semen, etc.;*
      - *Serious physical harm, including visible wounds*
      - *Gratuitous gore: any media that depicts excessively graphic or gruesome content related to death, violence or severe physical harm that is shared for sadistic purposes.*

One potential concern for enforcing policies around self-harm lies in driving members of self-harm communities off mainstream platforms such as Twitter toward "dark web" communities that are harder to monitor. Although this may be a valid concern, doing so is still likely to disrupt the ability of such communities to spread via social media contagion. This is because accessing and organizing on the "dark web" is much more difficult than doing so on mainstream platforms such as Twitter, and we speculate that relatively few teenagers know how to do this. Even for those who do know how, the challenge may be a sufficient obstacle for many to decide that the effort is not worth it. Therefore, disrupting these communities on mainstream platforms is likely to reduce the spread and contagion of self-harm.

[8] https://help.twitter.com/en/rules-and-policies/glorifying-self-harm
[9] https://help.twitter.com/en/rules-and-policies/media-policy

Conclusion

How many of these users have violated Twitter's own terms of service? The answer is vastly greater than zero—certainly in the thousands, and possibly in the hundreds of thousands.

Twitter currently requires users to be 13 or older. However, there is no enforcement mechanism to ensure that a user who claims to be 13 is, in fact, 13. Furthermore, much graphic content *permitted* on the platform is entirely inappropriate for 13-17 year olds, including "adult content" (which is often posted with self-harm hashtags in order for it to appear in the online self-harm communities), and material that is "excessively gory" or "violent."

But even *impermissible* material that is inappropriate for children according to Twitter's terms of service is in reality widely available, and is being consumed by children with psychiatric disorders whose disordered thinking and dangerous behavior is exacerbated by engaging with such content.

It is also a distinct possibility that as a result of an algorithm, young people seeking help to stop harming themselves could find themselves, instead, exposed to communities that encourage and celebrate their compulsion to cut themselves.

Twitter hosts a similar community of young people with anorexia, gamifying another life-threatening disorder. Like "Self-Harm Twitter," "Eating Disorder Twitter" (#edtwt) uses code for levels of severity. "Thinspo," for example, refers to being extremely underweight, but less so than "bonespo," for which the anorexic is so severely underweight that bones are clearly visible through the skin. Like children who find praise, encouragement, and instruction for slicing into their flesh with knives and razors, anorexics find a similar community on Twitter to reinforce their disorder. Both disorders can be life-threatening.

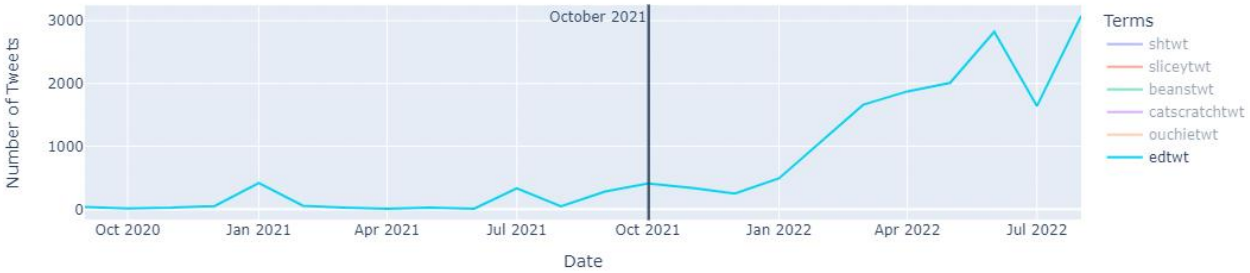


Fig 8. Timeline analysis for tweets and retweets for "edtwt," associated with "eating disorder twitter." Post associated with this hashtag are typically encouraging self-harm behaviors.

There are several reasons why policy teams at Twitter might fail to prioritize self-harm as a consequential community when it comes to policy enforcement. To begin with, the platform seems largely focused on moderating or banning content that upsets the Twitter community, often adjacent to political concerns, and this focus may help explain why this sort of content has slipped through the cracks. Because members of the self-harm community are not hostile and

celebrate rather than deride one another, members are unlikely to denounce or report one another to Twitter. Instead, members are affirming and encourage one another to cut more, cut more deeply, and to post images. If these networks continue to grow unabated on Twitter, so will the risk of increasingly severe or even fatal injuries.

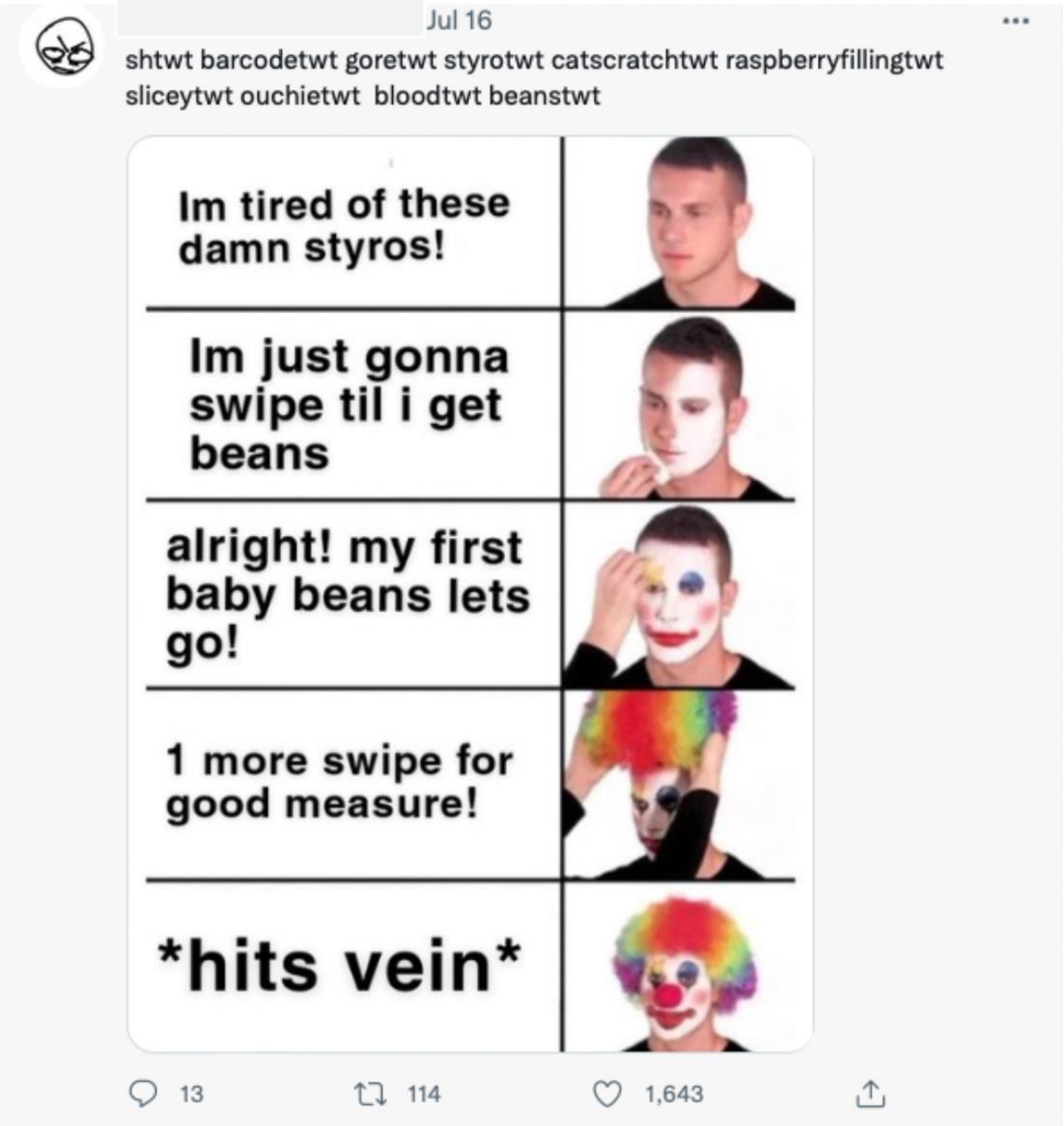
If children who self-harm (whether by cutting or by starving themselves) continue to find encouragement for increasing the severity of their injuries on Twitter rather than resources for getting help, Twitter will be an ongoing, potent accelerator for serious disorders. Twitter's inability to keep up with the evolution of coded language allows social contagions of self-harm to escape detection, metastasize, and persist on the platform. If Twitter and other platforms prove unable to enforce their own policies regarding suicide and self-harm, perhaps they should collectively create and fund an independent nonprofit social media platform auditor to assist them in protecting children.

Twitter as it is currently configured is not an appropriate platform for anyone under the age of 18.

Appendix

Examples of Suicide and Self Harm Policy Violations

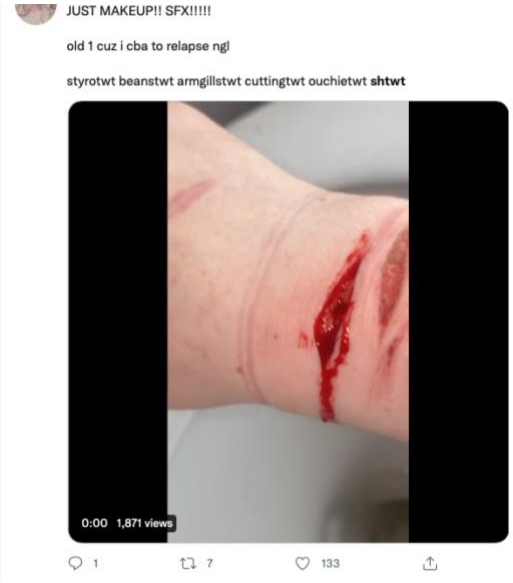

Aug 2
Replying to
What do you use for this? and what technique? I'm struggling and can't get that deep/:
3
11
Aug 3
pressure and fast
4
shtwt how are we
might just let my
Razors win tonight
20
250
1,581



Examples of Sensitive Media Policy Violations (Tweets below posted in August 2022.)

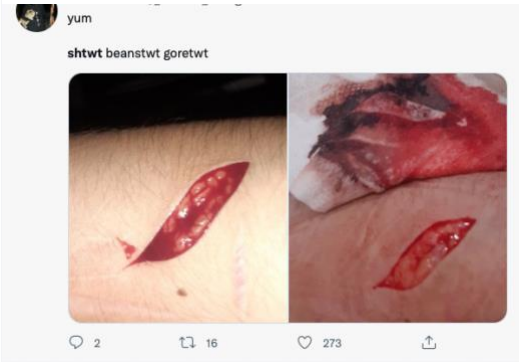

‼️TW-Sh mention/pictures‼️
slit myself again 🔪
getting stiches was so painful, I almost fell unconscious 😫
anyways guys, here are some lovely pics I took🤪🙏🏳️‍🌈✨🤪
#sliceytwt #shtwt #beanstwt #goretwt
1
1
14


Yourmom @gaymothefucker · Aug 22
‼️ TW - Selfharm pictures/mention ‼️
dw twitter, it's all sfx😋🙏😊
this looks weird as shit 🔪
i hate getting stitches guys💔
#shtwt #sliceytwt #goretwt #stitches #emofag
1
4


JUST MAKEUP!! SFX!!!!!
old 1 cuz i cba to relapse ngl
styrotwt beanstwt armgillstwt cuttingtwt ouchietwt shtwt
0:00 1,871 views
1
7
133


shtwt self harm a.m babycut sytro babysytro beans edtwt
5
13
225


Tw // selfharm
This is how I live ; we live ;
Shtwt sliceytwt ouchietwt barcodetwt
5
63


yum
shtwt beanstwt goretwt
2
16
273

Other Concerning Content

User asking followers in the shtwt and edtwt if they have ever thought about “intentionally hurting or killing a child.” The post received nearly 8,000 responses. 42% of respondents said yes. A majority of users responded yes to questions asking if users have thought about “intentionally killing a stranger” or “intentionally killing someone close to you.”

User emphasizing how shtwt provides a sense of community.