PX0304

The Cambridge Analytica Files Cambridge Analytica

This article is more than **6 years old**

Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters

'I made Steve Bannon's psychological warfare tool': meet the data war whistleblower

Mark Zuckerberg breaks silence on Cambridge Analytica

Cambridge Analytica whistleblower: 'We spent $1m harvesting millions of Facebook profiles' – video

Carole Cadwalladr *and* Emma Graham-Harrison

Sat 17 Mar 2018 18.03 EDT

The data analytics firm that worked with Donald Trump's election team and the winning Brexit campaign harvested millions of Facebook profiles of US voters, in one of the tech giant's biggest ever data breaches, and used them to build a powerful software program to predict and influence choices at the ballot box.

A whistleblower has revealed to the *Observer* how Cambridge Analytica - a company owned by the hedge fund billionaire Robert Mercer, and headed at the time by Trump's key adviser Steve Bannon - used personal information taken without authorisation in early 2014 to build a system that could profile individual US voters, in order to target them with personalised political advertisements.

Christopher Wylie, who worked with a Cambridge University academic to obtain the data, told the *Observer*: "We exploited Facebook to harvest millions of people's profiles. And built models to exploit what we knew about them and target their inner demons. That was the basis the entire company was built on."

Documents seen by the *Observer*, and confirmed by a Facebook statement, show that by late 2015 the company had found out that information had been harvested on an unprecedented scale. However, at the time it failed to alert users and took only limited steps to recover and secure the private information of more than 50 million individuals.

Profile
Cambridge Analytica: the key players

Show

The *New York Times* is reporting that copies of the data harvested for Cambridge Analytica could still be found online; its reporting team had viewed some of the raw data.

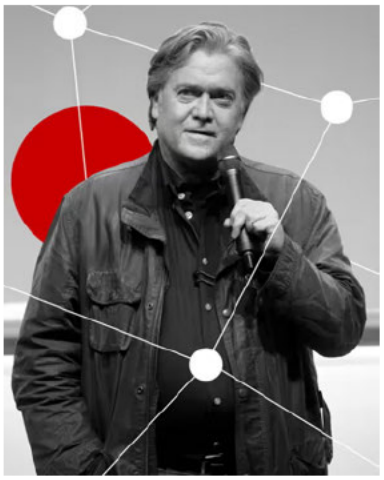
The data was collected through an app called thisisyourdigitallife, built by academic Aleksandr Kogan, separately from his work at Cambridge University. Through his company Global Science Research (GSR), in collaboration with Cambridge Analytica, hundreds of thousands of users were paid to take a personality test and agreed to have their data collected for academic use.

However, the app also collected the information of the test-takers' Facebook friends, leading to the accumulation of a data pool tens of millions-strong. Facebook's "platform policy" allowed only collection of friends' data to improve user experience in the app and barred it being sold on or used for advertising. The discovery of the unprecedented data harvesting, and the use to which it was put, raises urgent new questions about Facebook's role in targeting voters in the US presidential election. It comes only weeks after indictments of 13 Russians by the special counsel Robert Mueller which stated they had used the platform to perpetrate "information warfare" against the US.

Cambridge Analytica and Facebook are one focus of an inquiry into data and politics by the British Information Commissioner's Office. Separately, the Electoral Commission is also investigating what role Cambridge Analytica played in the EU referendum.

"We are investigating the circumstances in which Facebook data may have been illegally acquired and used," said the information commissioner Elizabeth Denham. "It's part of our ongoing investigation into the use of data analytics for political purposes which was launched to consider how political parties and campaigns, data analytics companies and social media platforms in the UK are using and analysing people's personal information to micro-target voters."

On Friday, four days after the *Observer* sought comment for this story, but more than two years after the data breach was first reported, Facebook announced that it was suspending Cambridge Analytica and Kogan from the platform, pending further information over misuse of data. Separately, Facebook's external lawyers warned the *Observer* it was making "false and defamatory" allegations, and reserved Facebook's legal position.



Key Trump adviser Steve Bannon Photograph: Alain Robert/Sipa/Rex/Shutterstock

The revelations provoked widespread outrage. The Massachusetts Attorney General Maura Healey announced that the state would be launching an investigation. "Residents deserve answers immediately from Facebook and Cambridge Analytica," she said on Twitter.

The Democratic senator Mark Warner said the harvesting of data on such a vast scale for political targeting underlined the need for Congress to improve controls. He has proposed an Honest Ads Act to regulate online political advertising the same way as television, radio and print. "This story is more evidence that the online political advertising market is essentially the Wild West. Whether

it's allowing Russians to purchase political ads, or extensive micro-targeting based on ill-gotten user data, it's clear that, left unregulated, this market will continue to be prone to deception and lacking in transparency," he said.

Last month both Facebook and the CEO of Cambridge Analytica, Alexander Nix, told a parliamentary inquiry on fake news: that the company did not have or use private Facebook data.

Simon Milner, Facebook's UK policy director, when asked if Cambridge Analytica had Facebook data, told MPs: "They may have lots of data but it will not be Facebook user data. It may be data about people who are on Facebook that they have gathered themselves, but it is not data that we have provided."

Cambridge Analytica's chief executive, Alexander Nix, told the inquiry: "We do not work with Facebook data and we do not have Facebook data."

Wylie, a Canadian data analytics expert who worked with Cambridge Analytica and Kogan to devise and implement the scheme, showed a dossier of evidence about the data misuse to the *Observer* which appears to raise questions about their testimony. He has passed it to the National Crime Agency's cybercrime unit and the Information Commissioner's Office. It includes emails, invoices, contracts and bank transfers that reveal more than 50 million profiles - mostly belonging to registered US voters - were harvested from the site in one of the largest-ever breaches of Facebook data. Facebook on Friday said that it was also suspending Wylie from accessing the platform while it carried out its investigation, despite his role as a whistleblower.

At the time of the data breach, Wylie was a Cambridge Analytica employee, but Facebook described him as working for Eunoia Technologies, a firm he set up on his own after leaving his former employer in late 2014.

The evidence Wylie supplied to UK and US authorities includes a letter from Facebook's own lawyers sent to him in August 2016, asking him to destroy any data he held that had been collected by GSR, the company set up by Kogan to harvest the profiles.

That legal letter was sent several months after the *Guardian* first reported the breach and days before it was officially announced that Bannon was taking over as campaign manager for Trump and bringing Cambridge Analytica with him.

Cambridge Analytica interactive

"Because this data was obtained and used without permission, and because GSR was not authorised to share or sell it to you, it cannot be used legitimately in the future and must be deleted immediately," the letter said.

Facebook did not pursue a response when the letter initially went unanswered for weeks because Wylie was travelling, nor did it follow up with forensic checks on his computers or storage, he said.

"That to me was the most astonishing thing. They waited two years and did absolutely nothing to check that the data was deleted. All they asked me to do was tick a box on a form and post it back."

Paul-Olivier Dehaye, a data protection specialist, who spearheaded the investigative efforts into the tech giant, said: "Facebook has denied and denied and denied this. It has misled MPs and congressional investigators and it's failed in its duties to respect the law.

"It has a legal obligation to inform regulators and individuals about this data breach, and it hasn't. It's failed time and time again to be open and transparent."

> ***We exploited Facebook to harvest millions of profiles. And built models to exploit that and target their inner demons***
> **Christopher Wylie**

A majority of American states have laws requiring notification in some cases of data breach, including California, where Facebook is based.

Facebook denies that the harvesting of tens of millions of profiles by GSR and Cambridge Analytica was a data breach. It said in a statement that Kogan "gained access to this information in a legitimate way and through the proper channels" but "did not subsequently abide by our rules" because he passed the information on to third parties.

Facebook said it removed the app in 2015 and required certification from everyone with copies that the data had been destroyed, although the letter to Wylie did not arrive until the second half of 2016. "We are committed to vigorously enforcing our policies to protect people's information. We will take whatever steps are required to see that this happens," Paul Grewal, Facebook's vice-president, said in a statement. The company is now investigating reports that not all data had been deleted.

Kogan, who has previously unreported links to a Russian university and took Russian grants for research, had a licence from Facebook to collect profile data, but it was for research purposes only. So when he hoovered up information for the commercial venture, he was violating the company's terms. Kogan maintains everything he did was legal, and says he had a "close working relationship" with Facebook, which had granted him permission for his apps.

Quick Guide
How the Cambridge Analytica story unfolded

Show

The *Observer* has seen a contract dated 4 June 2014, which confirms SCL, an affiliate of Cambridge Analytica, entered into a commercial arrangement with GSR, entirely premised on harvesting and processing Facebook data. Cambridge Analytica spent nearly $1m on data collection, which yielded more than 50 million individual profiles that could be matched to electoral rolls. It then used the test results and Facebook data to build an algorithm that could analyse individual Facebook profiles and determine personality traits linked to voting behaviour.

The algorithm and database together made a powerful political tool. It allowed a campaign to identify possible swing voters and craft messages more likely to resonate.

"The ultimate product of the training set is creating a 'gold standard' of understanding personality from Facebook profile information," the contract specifies. It promises to create a database of 2 million "matched" profiles, identifiable and tied to electoral registers, across 11 states, but with room to expand much further.

Key players

At the time, more than 50 million profiles represented around a third of active North American Facebook users, and nearly a quarter of potential US voters. Yet when asked by MPs if any of his firm's data had come from GSR, Nix said: "We had a relationship with GSR. They did some research for us back in 2014. That research proved to be fruitless and so the answer is no."

Cambridge Analytica said that its contract with GSR stipulated that Kogan should seek informed consent for data collection and it had no reason to believe he would not.

GSR was "led by a seemingly reputable academic at an internationally renowned institution who made explicit contractual commitments to us regarding its legal authority to license data to SCL Elections", a company spokesman said.

SCL Elections, an affiliate, worked with Facebook over the period to ensure it was satisfied no terms had been "knowingly breached" and provided a signed statement that all data and derivatives had been deleted, he said. Cambridge Analytica also said none of the data was used in the 2016 presidential election.

Steve Bannon's lawyer said he had no comment because his client "knows nothing about the claims being asserted". He added: "The first Mr Bannon heard of these reports was from media inquiries in the past few days." He directed inquires to Nix.

The Observer

The Observer is the world's oldest Sunday newspaper, founded in 1791. It is published by Guardian News & Media and is editorially independent.

More on this story

Hackers steal customer data from Europe's largest

Greater Manchester police officers' data hacked in

Met police on high alert after supplier IT security

PX0305

Meta

Back to Newsroom

Meta

An Update on Our Plans to Restrict Data Access on Facebook

April 4, 2018

> "We believe these changes will better protect people's information while still enabling developers to create useful experiences."

By Mike Schroepfer, Chief Technology Officer

Two weeks ago we promised to take a hard look at the information apps can use when you connect them to Facebook as well as other data practices. Today, we want to update you on the changes we're making to better protect your Facebook

information. We expect to make more changes over the coming months — and will keep you updated on our progress. Here are the details of the nine most important changes we are making.

Events API: Until today, people could grant an app permission to get information about events they host or attend, including private events. This made it easy to add Facebook Events to calendar, ticketing or other apps. But Facebook Events have information about other people's attendance as well as posts on the event wall, so it's important that we ensure apps use their access appropriately. Starting today, apps using the API will no longer be able to access the guest list or posts on the event wall. And in the future, only apps we approve that agree to strict requirements will be allowed to use the Events API.

Groups API: Currently apps need the permission of a group admin or member to access group content for closed groups, and the permission of an admin for secret groups. These apps help admins do things like easily post and respond to content in their groups. However, there is information about people and conversations in groups that we want to make sure is better protected. Going forward, all third-party apps using the Groups API will need approval from Facebook and an admin to ensure they benefit the group. Apps will no longer be able to access the member list of a group. Apps can see all posts and comments in the group, but they can't see a group member's name, profile picture or that the group member is the author of posts and comments unless the member allows access. ***(Updated on March 2, 2022 at 12:00PM PT to clarify what information apps can see.)***

Pages API: Until today, any app could use the Pages API to read posts or comments from any Page. This let developers create tools for Page owners to help them do things like schedule posts and reply to comments or messages. But it also let apps access more data than necessary. We want to make sure Page information is only available to apps providing useful services to our community. So starting today, all future access to the Pages API will need to be approved by Facebook.

Facebook Login: Two weeks ago we announced important changes to Facebook Login. Starting today, Facebook will need to approve all apps that request access to information such as check-ins, likes, photos, posts, videos, events and groups. We started approving these permissions in 2014, but now we're tightening our review process — requiring these apps to agree to strict requirements before they can

access this data. We will also no longer allow apps to ask for access to personal information such as religious or political views, relationship status and details, custom friends lists, education and work history, fitness activity, book reading activity, music listening activity, news reading, video watch activity, and games activity. In the next week, we will remove a developer's ability to request data people shared with them if it appears they have not used the app in the last 3 months.

Instagram Platform API: We're making the recently announced deprecation of the Instagram Platform API effective today. You can find more information here.

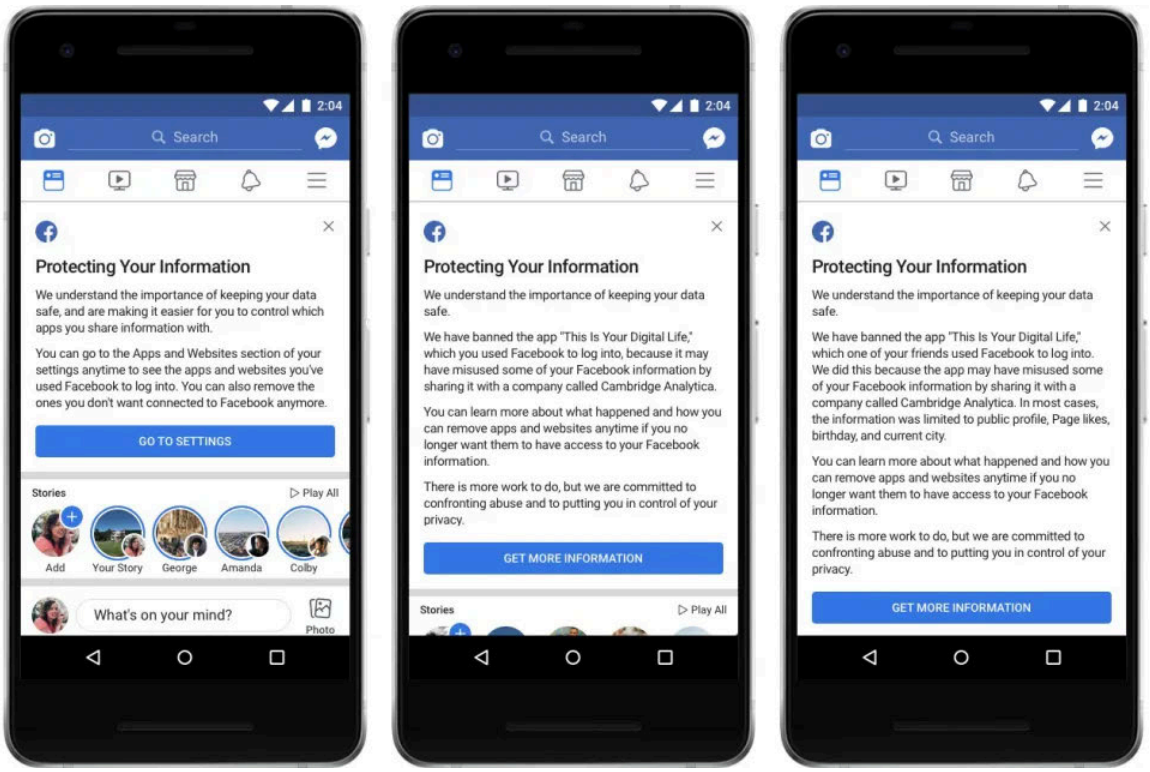
Search and Account Recovery: Until today, people could enter another person's phone number or email address into Facebook search to help find them. This has been especially useful for finding your friends in languages which take more effort to type out a full name, or where many people have the same name. In Bangladesh, for example, this feature makes up 7% of all searches. However, malicious actors have also abused these features to scrape public profile information by submitting phone numbers or email addresses they already have through search and account recovery. Given the scale and sophistication of the activity we've seen, we believe most people on Facebook could have had their public profile scraped in this way. So we have now disabled this feature. We're also making changes to account recovery to reduce the risk of scraping as well.

Call and Text History: Call and text history is part of an opt-in feature for people using Messenger or Facebook Lite on Android. This means we can surface the people you most frequently connect with at the top of your contact list. We've reviewed this feature to confirm that Facebook does not collect the content of messages — and will delete all logs older than one year. In the future, the client will only upload to our servers the information needed to offer this feature — not broader data such as the time of calls.

Data Providers and Partner Categories: Last week we announced our plans to shut down Partner Categories, a product that lets third-party data providers offer their targeting directly on Facebook.

App Controls: Finally, starting on Monday, April 9, we'll show people a link at the top of their News Feed so they can see what apps they use — and the information they have shared with those apps. People will also be able to remove apps that they no

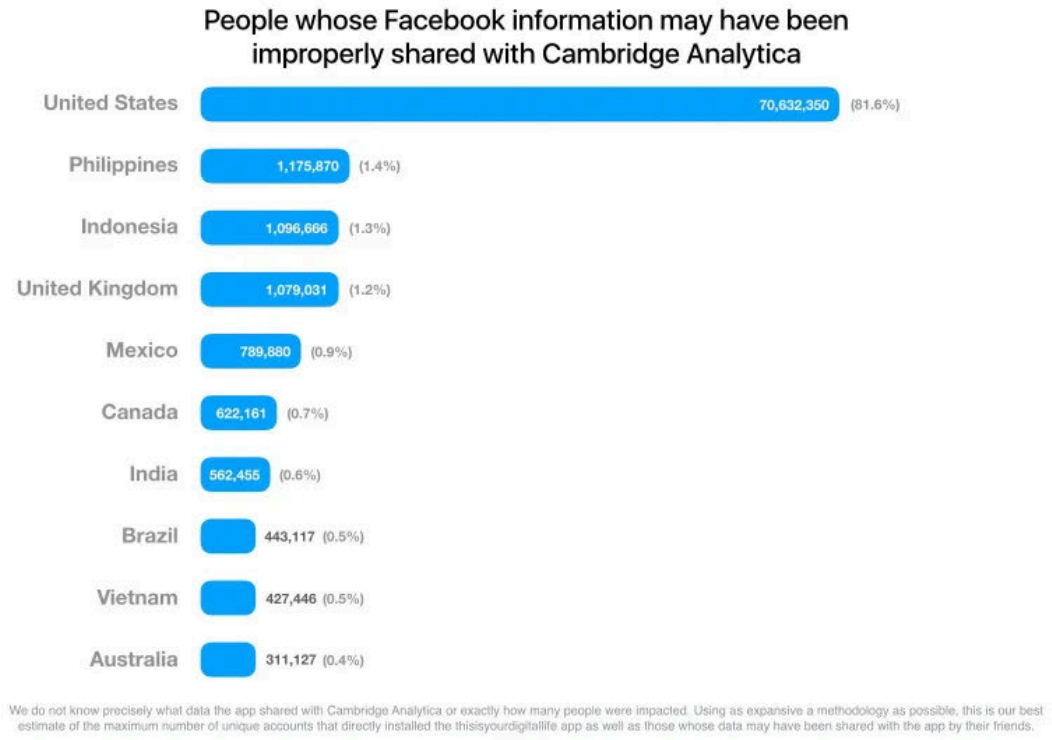
longer want. As part of this process we will also tell people if their information may have been improperly shared with Cambridge Analytica.



Updated April 9, 2018: Three versions of the messages we're sending to people based on whether they've been affected by the app "This Is Your Digital Life." These messages link to facebook.com/help/yourinfo.

In total, we believe the Facebook information of up to 87 million people — mostly in the US — may have been improperly shared with Cambridge Analytica.

https://about.fb.com/news/2018/04/restricting-data-access/



(***Update on May 1, 2018:*** Click here to see a state-by-state breakdown in the US of people whose Facebook information may have been improperly shared with Cambridge Analytica.)

Overall, we believe these changes will better protect people's information while still enabling developers to create useful experiences. We know we have more work to do — and we'll keep you updated as we make more changes. You can find more details on the platform changes in our Facebook Developer Blog.

Downloads

Cambridge Analytica Graph

App Controls Screenshots

Categories:
Company News, Data and Privacy, Meta, Product News

Tags: Platform, Security News

RELATED NEWS

Meta

Increasing Our Ads Transparency

We're giving you more information about why you see the ads you do on our apps.

February 14, 2023

Topics

Company News

Technology and Innovation

Data and Privacy

Safety and Expression

Combating Misinformation

Economic Opportunity

Election Integrity

Strengthening Communities

Diversity and Inclusion

PX0307

Calm down. Breathe. We hear you.

by Mark Zuckerberg 10:45pm Tuesday, Sep 5

We've been getting a lot of feedback about Mini-Feed and News Feed. We think they are great products, but we know that many of you are not immediate fans, and have found them overwhelming and cluttered. Other people are concerned that non-friends can see too much about them. We are listening to all your suggestions about how to improve the product; it's brand new and still evolving.

We're not oblivious of the Facebook groups popping up about this (by the way, Ruchi is not the devil). And we agree, stalking isn't cool; but being able to know what's going on in your friends' lives is. This is information people used to dig for on a daily basis, nicely reorganized and summarized so people can learn about the people they care about. You don't miss the photo album about your friend's trip to Nepal. Maybe if your friends are all going to a party, you want to know so you can go too. Facebook is about real connections to actual friends, so the stories coming in are of interest to the people receiving them, since they are significant to the person creating them.

We didn't take away any privacy options. [Your privacy options remain the same.] The privacy rules haven't changed. None of your information is visible to anyone who couldn't see it before the changes. If you turned off your wall to non-friends, no one who is not your friend will be able to see a post on your wall. Your friends can still see it; it hasn't changed. Secret groups and secret events remain secret from other people. Pokes and messages remain as private interactions. Nothing you do is being broadcast; rather, it is being shared with people who care about what you do—your friends.

We're going to continue to improve Facebook, and we want you to be part of that process. Test out the products and continue to provide us feedback. Use your privacy settings so you can feel most comfortable using the site.

We hear you, and we appreciate the feedback.

Stay tuned... Mark

To post a comment, you must be logged in. To register for Facebook, click here.

Monthly Archive:
October (3)
September (7)
August (4)

Subscribe to this Blog:
Facebook Blog (Atom)
Learn More about Atom »

Facebook © 2006

PX0308

The Future of Facebook

Laura Locke

July 17, 2007

Paul Sakuma / AP

Facebook.com's mastermind Mark Zuckerberg

In his first interview with TIME, Facebook CEO Mark Zuckerberg sat down with reporter Laura Locke to talk about Facebook's rapid growth spurt, IPO rumors, future plans and the pressures of being a 23-year-old CEO in Silicon Valley.

TIME: Facebook is undergoing a huge period of growth. With more than 150,000 new users signing up daily, it is growing three times as fast as rival MySpace. What do you attribute that spike to?

Zuckerberg: For a while we actually constrained our growth. We made it so that only people in college could sign up. Initially it was only available to people at Harvard, where I was at college. We rolled it out to all the colleges, all the high schools, then a bunch of companies could sign up, and now everyone can sign up. It may seem like the growth is really accelerating at a crazy rate, but it's actually been growing and doubling about once every six months for quite a while.

TIME: Is Facebook's popularity connected to its focus on authenticity? On your site, misrepresentation of your real self is a violation of company policy.

Zuckerberg: That's the critical part of it. Our whole theory is that people have real connections in the world. People communicate most naturally and effectively with their friends and the people around them. What we figured is that if we could model what those connections were, [we could] provide that information to a set of applications through which people want to share information, photos or videos or events. But that only works if those relationships are real. That's a really big difference between Facebook and a lot of other sites. We're not thinking about ourselves as a community — we're not trying to build a community — we're not trying to make new connections.

TIME: Why do you describe Facebook as a "social utility" rather than a "social network?"

Zuckerberg: I think there's confusion around what the point of social networks is. A lot of different companies characterized as social networks have different goals — some serve the function of business networking, some are media portals. What we're trying to do is just make it really efficient for people

to communicate, get information and share information. We always try to emphasize the utility component.

TIME: In September you rebuffed Yahoo's offer to buy Facebook for nearly $1 billion. Before that, Viacom put up a $750 million bid. And about two months ago you clearly said Facebook would stay independent. Is that still the plan?

Zuckerberg: That has always been the plan. As a company we're very focused on what we're building and not as focused on the exit. We just believe that we're adding a certain amount of value to people's lives if we build a very good product. That's the reason why more than half of our users use the product every day — it's a more efficient way for them to communicate with their friends and get information about the people around them than anything else they can do. We're not really looking to sell the company. We're not looking to IPO anytime soon. It's just not the core focus of the company.

TIME: So, if Facebook isn't for sale and there's no IPO in the works, how do you intend to satisfy your investors who put a total of $38 million into the company?

Zuckerberg: Well, they're actually really supportive of this. What they want is to build a really great company, too. And if you think about the timeframe over which this has happened — we took our venture round from Accel Partners just about two years ago — they're not in a rush. We have plenty of time to build something good.

TIME: Facebook is looking to hire a stock administrator, isn't that a signal you're preparing for an IPO?

Zuckerberg: Well, no. [Pause.] I mean, we grant options to all of our employees. At this point we have more than 250. It's a core part of compensation, so you want to make sure you get it right for people. At some point in the future, if we get a chance to go down that [IPO] path, it will be valuable to have that — it's a part of building out the company. I think it's funny that people are paying so much attention to that.

TIME: The frenzy surrounding Facebook seems to have intensified quite dramatically over the past several months. What do you think is behind the company's newfound cachet?

Zuckerberg: I think the most recent surge, at least in the press, is around the launch of Facebook Platform. For the first time we're allowing developers who don't work at Facebook to develop applications just as if they were. That's a big deal because it means that all developers have a new way of doing business if they choose to take advantage of it. There are whole companies that are forming whose only product is a Facebook Platform application. That provides an opportunity for them, it provides an opportunity for people who want to make money by investing in those companies, and I think that's something that's pretty exciting to the business community. It's also really exciting to our users because it means that a whole new variety of services are going to be made available.

TIME: What's your grand plan for the company? How do you see it evolving over the next three to five years?

Zuckerberg: It's tough to say, exactly, what things will look like in three to five years, but there's a lot of work to do in just moving along the path that we've already set out. Right now we have 30 million active users on Facebook. There's a lot more to go. And there a lot of different applications that are going to be developed to allow people to share information in different ways. I would expect the user base will grow [and there will be] more ways for advertisers to reach people and communicate in a very natural way, just like users communicate with each other. All these things will just get more and more evolved.

TIME: Beyond Facebook's exclusive advertising deal with Microsoft, which gives the software giant the right to sell ads on the site, what are some of your ideas about monetizing your 30 million users?

Zuckerberg: Advertising works most effectively when it's in line with what people are already trying to do. And people are trying to communicate in a certain way on Facebook — they share information with their friends, they learn about what their friends are doing —so there's really a whole new opportunity for a new type of advertising model within that. And I think we'll see more in the next couple months or years on that.

TIME: With more than 40 billion page views every month, Facebook is the sixth most trafficked site in the U.S., and the top photo-sharing site. What are your international expansion plans?

Zuckerberg: Right now a lot of our growth is happening internationally. We have more than 10% or 15% of the population of Canada on the site. The U.K. has a huge user base. We haven't translated the site yet, but that's something we're working on and it should be done soon. What we're doing is pretty broadly applicable to people in all different age groups and demographics and places around the world.

TIME: You recently took off for a summer vacation, what did you do?

Zuckerberg: Hang out with my family.

TIME: What's a typical day like for the guy who founded Facebook in his Harvard dorm room just three years ago before becoming a full-time entrepreneur?

Zuckerberg: I wake up in the morning, I walk to work because I live four blocks from one of our offices, and I work, meet with people, and discuss things all day, and then I go home and go to sleep. I don't have an alarm clock. If someone needs to wake me up, then I have my BlackBerry next to me.

TIME: You're a 23-year-old Silicon Valley CEO. How do you deal with all pressures that come along with running a hyper-fast paced, high-profile technology company?

Zuckerberg: I was watching an interview with Steve Jobs the other day, in which he said that 'In order to be doing something like this, you have to really, really like what you're doing, because otherwise it just doesn't make sense.' The demands and the amount of work that it takes to put something like [Facebook] into place, it's just so much that if you weren't completely into what you were doing and you didn't think it was an important thing, then it would be irrational to spend that much time on it. Part of the reason why this is fun is because we've managed to build a team of really smart people who come from different backgrounds and have different experiences and think in different ways. People constantly try to put us in a bucket: are we trying to sell the company? What are we trying to do? What is the business strategy? People are often more interested in why we're hiring a stock-options administrator. Whereas for me and a lot of people around me, that's not really what we focus on. We're just focused on building things.

PX0310

STEVEN LEVY BUSINESS OCT 6, 2022 9:00 AM

Meta Rethinks the Philosophy of the Facebook Feed

WIRED spoke with the head of Facebook about competition from TikTok and his theory that algorithms can enhance our personal relationships.

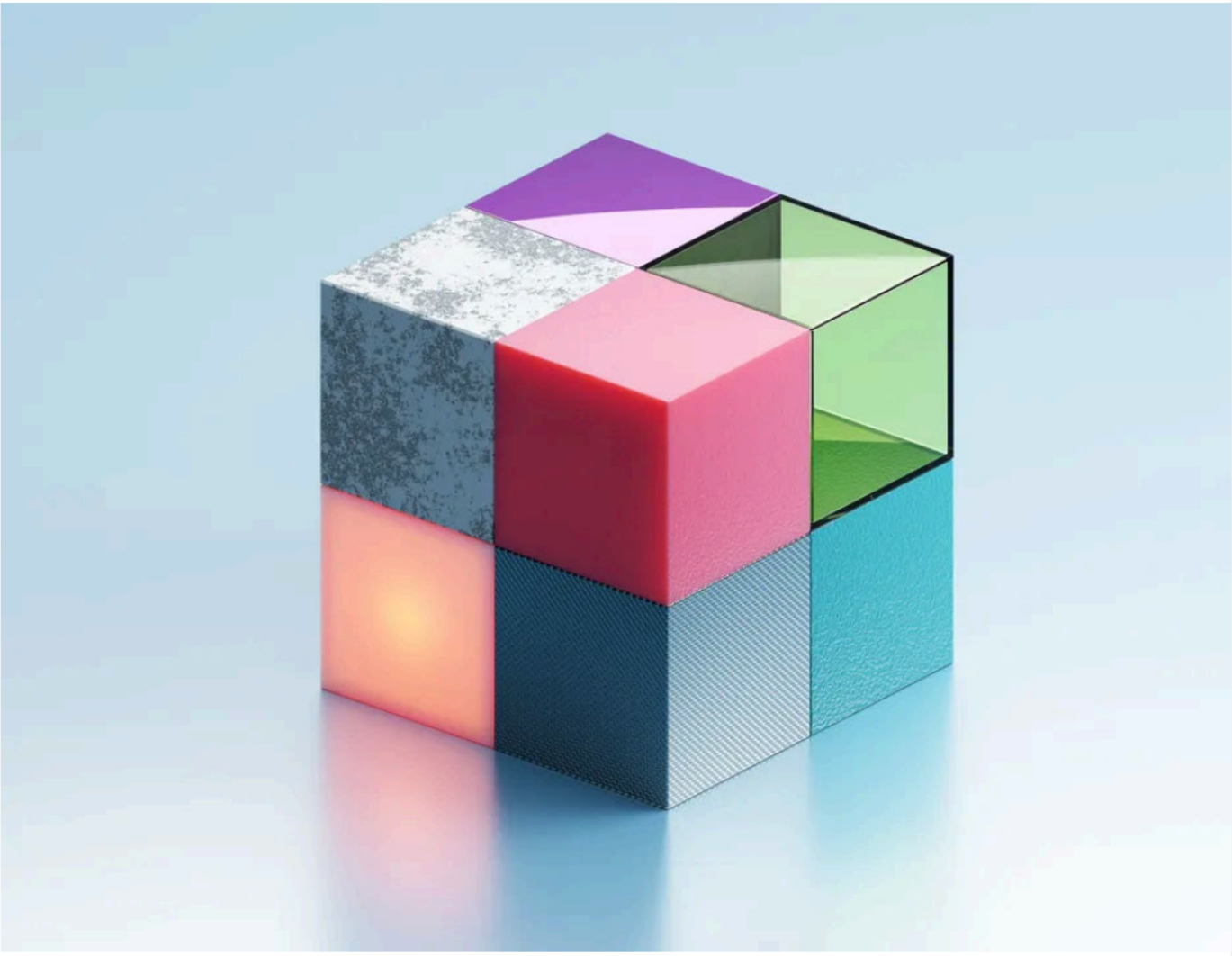


ILLUSTRATION: JORG GREUEL/GETTY IMAGES

SAVE

THE FACEBOOK APP, the blockbuster that powers the company we now call Meta, has been spending time in the juggernaut infirmary. While Feed is still the company's most popular product, its user base has stopped growing, in stark contrast to that of its fierce new competitor, TikTok. In response, Meta's leaders in July tweaked Facebook's iconic home page. An earlier leaked memo revealed that their intent was to transform Feed from a list of posts mostly from friends into a "discovery engine" of "unconnected content" that can come from anywhere—much like the addictive feed at the heart of TikTok.

Yesterday, Facebook announced further changes to its flagship Blue App that seem like a way for Facebook to hone its discovery engine's algorithms. Users will be able to tell Facebook if they want to see more or less of certain kinds of posts. The feedback will temporarily boost or suppress the ranking score for that kind of post. A second customization option can be used to specify how much you'd like to see posts from, respectively, friends and family, groups, and public figures. This feature is accessible from a page called Feed Preferences, which is only a moderate challenge to locate.

Tom Alison COURTESY OF META

As a longtime follower of the evolution of the feed—from the moment that Zuckerberg outlined its possibilities in his private notebook—I was curious about this shift in philosophy, one that will affect content still viewed by 3 billion people. I connected with Tom Alison, a 12-year veteran at the company who is head of Facebook. (He spent the first part of his tenure on the growth team, working on the notorious People You May Know feature.) The conversation has been edited for length and clarity.

Steven Levy: Tell me why you're making these changes to Feed.

Tom Alison: It's all about new ways that you can have agency and control over what you see in Feed and in Facebook. The first is what we call "see more, see less." On certain posts in your feed we offer a way to give us feedback—do you want to see more or see less of this type of content? You've always been able to hide somebody or block somebody, or put somebody in your favorites. But those are heavyweight,

long-lasting actions. "See more, see less" applies for the next 30 or 60 days. We're going to continue looking at likes, comments, and all the different ways we personalize content for you. But this change gives people a very direct way to tell us, "Hey, I want to see more of this type of content, or less." As we build out what we're calling the Discovery Engine, this is one of the ways that we're letting people tell us what they want from the experience.

Can you describe what the Discovery Engine really is?

The discovery engine is the home screen of Facebook. It's Stories from your friends, it's Reels from the creators and the friends that you follow. It includes both recommendations and content from people that you're connected to. The idea is that Facebook is the place where you discover some of the things that you have in common with others.

When I tell Facebook I want to see more or less of something, does that mean I'm telling it, for instance, that I don't want to see politics?

For now, the feature acts mostly on the author of the content. But we are looking at ways to also consider other signals and start asking, "OK, Steven, did you mean more content from Tom, or did you mean more content about the New York Yankees?"

The other feature you announced allows users to ask for more content from friends and family. I wonder if that's a corrective to the July announcement that threw open the home feed to all sorts of content from people you've never met. Isn't the Discovery Engine a departure from Facebook's original focus on people in your social graph?

Facebook is still at its core about friends and family. Our discovery engine involved a few different things. One is, how do we connect you with the people that you already know? But also, how do we connect you with the people that you might want to know.

You once worked on a product called People You May Know. Now Meta is telling me about people I *should* know, even if I'll never meet them. That's a big difference.

That's something people have told us that they really want to see, especially young adults. People want to use Facebook and these social media apps to explore more of their interests. I've chosen to be on some cooking groups on Facebook, like a great group where all the people share new restaurants in the Bay Area. I love it. But now Facebook is also learning to say, "Hey, there's these great cooking creators out there," and it's learned that I'm interested in particular in barbecuing. It's been showing me content from some great creators who are inspiring me to cook and to go follow them.

You mention young people. Facebook has been losing that demographic for years, and very few young people use it, or at least many claim they don't. Since you began changing Feed to specifically cater to this audience, have you seen an uptick in their use of the product?

When we say young, we mean young adults, 18- to 29-year-olds. We started really focusing our product development on them, around Reels, around recommendations, around making it easier for people to share to Stories or to messaging, in addition to Feed. All of that work is inspired by the research we built around the needs of young adults. We have a lot of work to do still, but I'm feeling pretty optimistic about what I'm seeing so far.

The cynical way to view your changes is that young people now go to TikTok over Facebook, not to make connections but just to entertain themselves. And you don't want to miss out on that.

People still want to share with their friends, but they're increasingly using things like our Stories product, which is ephemeral, where the story lasts for 24 hours. And they want to share with their friends over messaging experiences, where I can have a more intimate conversation with a few friends.
But even with this kind of broadcast posting [directed to a potential mass of people as opposed to those you know], people and friends connect over the stuff of life, right? People are talking about a new TV show or song, or playing *Wordle*. We're talking about that in our feeds together. Increasingly, that might happen in the context of a recommendation. I got a recommendation the other day for a band that makes techno music with a banjo. My wife and I really like that kind of music, and so I shared it with her over messaging. That was a very meaningful connection for her—

based on a recommendation that I got on Facebook from somebody that I didn't know.

We're still connecting you with your friends and the people that you care about, but increasingly I think a lot of these content inspirations are going to be the conversation starter for you and a friend. It doesn't necessarily have to be another friend in your network that starts the conversation.

How much time are people spending on Feed now, compared to a few years ago?

It's hard to say, but I'll tell you how attention is shifting at a high level. When people see friend content on Facebook, they're as likely to see it in a Story as they are to see it in Feed. That's why we've not only been making Stories easy to access and investing more in the creative tools, but we've started doing some work on integrating Stories content into Feeds, so that you don't miss it. Also, more than half the time spent on Facebook is now spent on video. We're putting Reels into Feed, and into recommendations, but we also have longer-form video and live video that continues to be part of the overall offering.

You've mentioned recommendations a number of times. In the past Facebook has been cited for recommending things that take people into dangerous rabbit holes. How are you protecting people from that?

We've been building integrity into everything we do for a long time now. There's community standards that govern what's allowed and not allowed on Facebook. We also have recommendation guidelines that we publish. Those are an even higher bar than our community standards. And we have very, very deep work on recommendations and AI systems, making sure that they're, that they're safe.

You Might Also Like ...

- Navigate election season with our WIRED Politics Lab **newsletter** and **podcast**
- She showed President Biden ChatGPT—and helped set the course for AI
- Laser weapons are finally poised to enter the battlefield

- What happens when a romance writer gets locked out of Google Docs
- **Give your back a break**: Here are the best office chairs we've tested

Steven Levy covers the gamut of tech subjects for WIRED, in print and online, and has been contributing to the magazine since its inception. His weekly column, Plaintext, is exclusive to subscribers online but the newsletter version is open to all—**sign up here**. He has been writing about technology for... Read more

EDITOR AT LARGE

TOPICS FACEBOOK SOCIAL MEDIA TIKTOK PRIVACY ALGORITHMS

READ MORE

BUSINESS

The Latest Online Culture War Is Humans vs. Algorithms

Ever feel like Instagram or TikTok algorithms know you a bit too well? The backlash against automated curation is building, and new algorithm-free platforms are springing up.

ELANA KLEIN

TECH

How TikTok Unleashed a Pressure Campaign on Congress

Users inundated lawmakers with calls after they were warned that Congress might enact a "total ban" on the app. Yet the bill at hand is showing no signs of slowing down.

CALEB ECARMA, VANITY FAIR

STYLE

William Hanson Is Emily Post for the TikTok Generation

Younger viewers are saying "yes, please" to Hanson's bite-size videos on manners and approach to teaching etiquette via social media.

ELIZA BROOKE, VANITY FAIR

DEATH VALLEY POSTCARD

The Death Valley Lake That's Gone in a Flash

Lake Manly forms in Badwater Basin only after especially heavy rains. Paddlers grab their paddles and go.

MEG BERNHARD, THE NEW YORKER

POLITICS

TikTok Is Facing Its Greatest Threat Yet

Today, the House of Representatives is set to vote on a bill that could result in a ban of the app in the US.

MAKENA KELLY

CULTURE

No One Knows What TikTok Is

This week, US lawmakers began a new push to ban TikTok. At the heart of it is a lot of confusion about what the app is and what it means.

ANGELA WATERCUTTER

POLITICS

The Kate Middleton Photo Controversy Is an Inexplicable Mess

Apparent manipulations to an official image of the Princess of Wales led wire services to issue a "kill notification" on the photo.

BRIAN BARRETT

BUSINESS

Why FOMO Is Fueling Nvidia Stock

The investment frenzy around AI is driving valuation so much that some investors are wondering when the bubble might burst.

NICK BILTON, VANITY FAIR

COOKIES SETTINGS

PX0311

Introducing Threads: A New Way to Share With Text

Update on April 2, 2024 at 5:00AM PT:

Our fact-checking program is now live on Threads. Fact-checking partners now have the ability to review and rate misinformation on Threads directly.

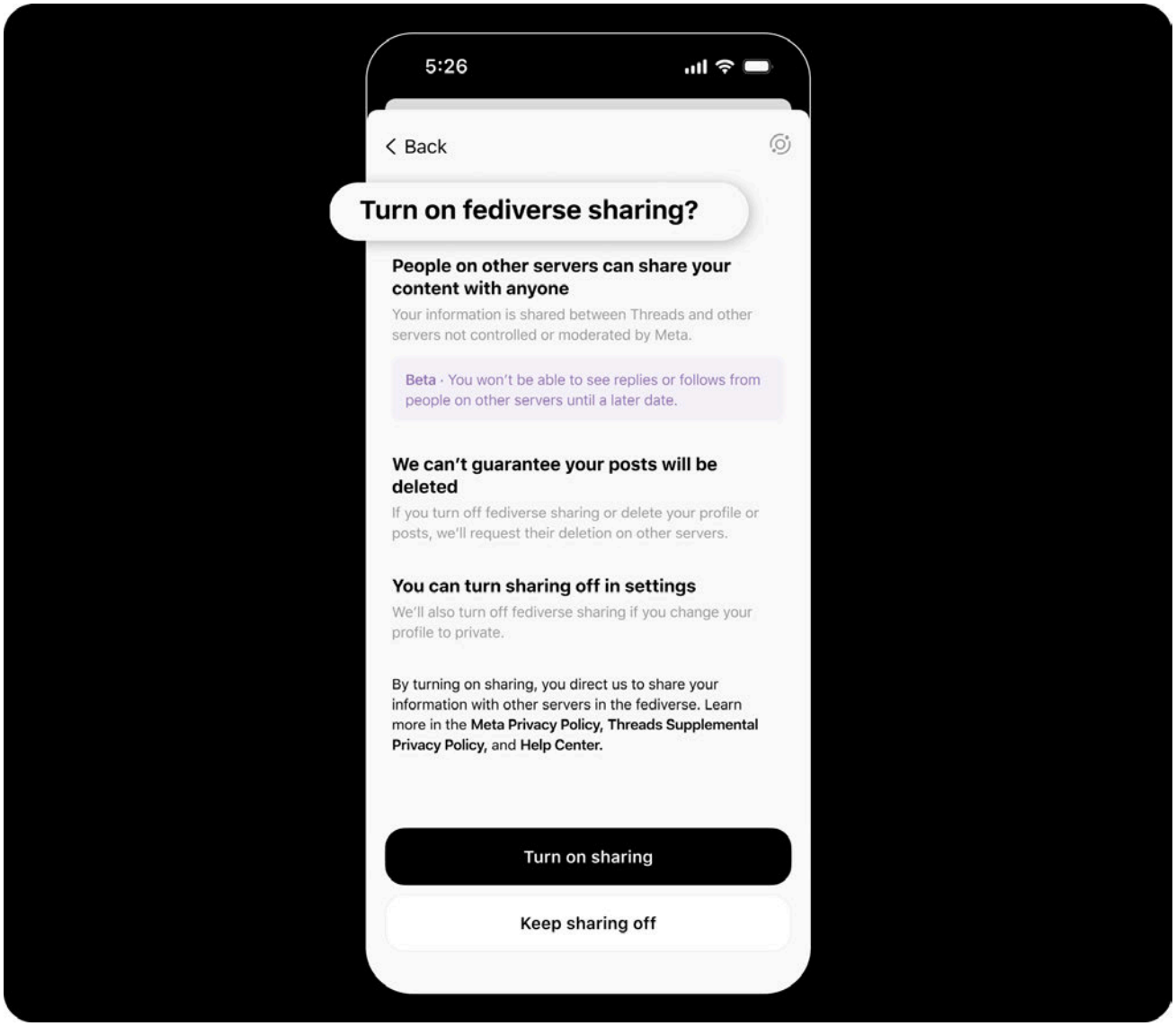
Update on March 21, 2024 at 11:00AM PT:

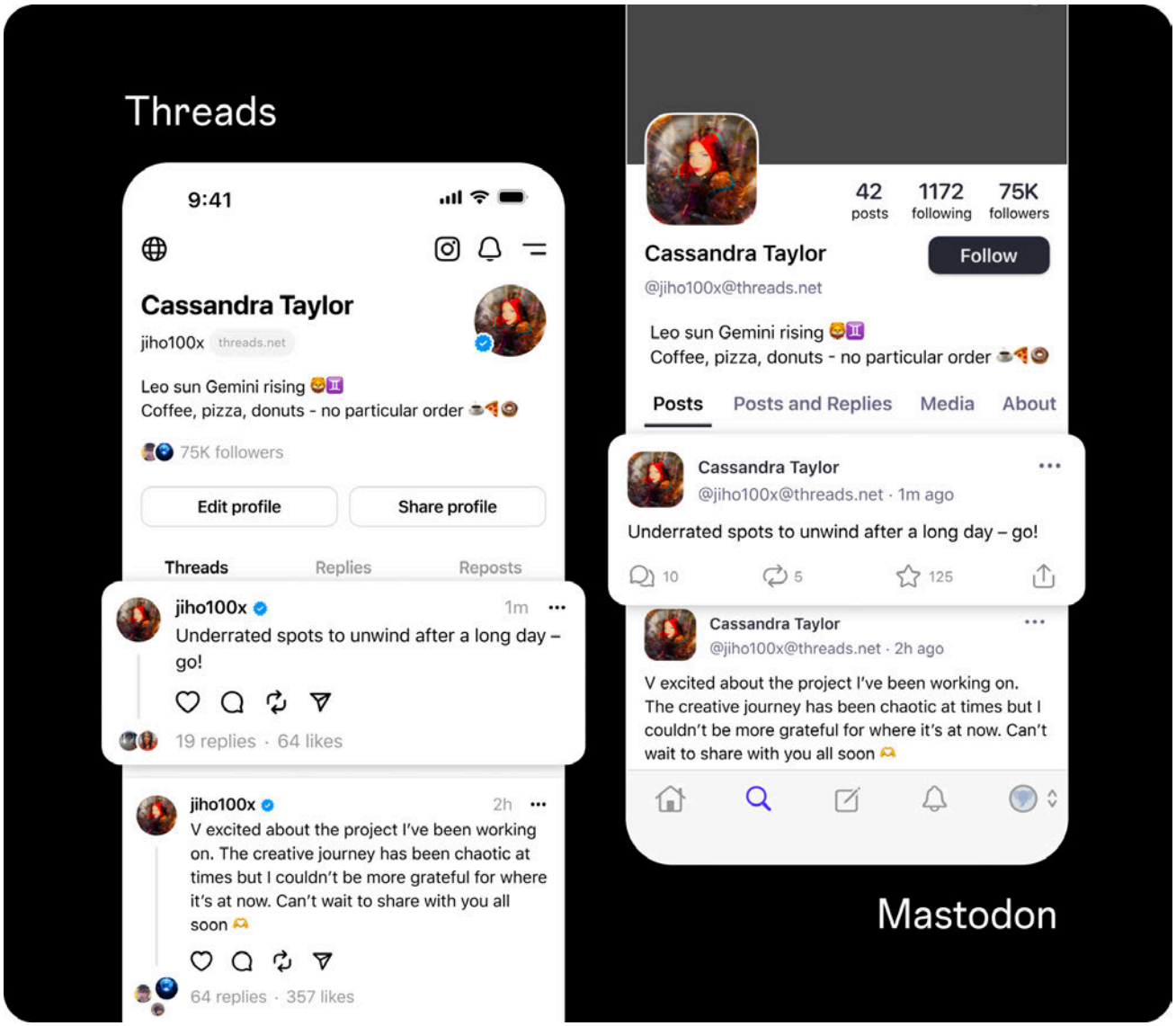
Mark Zuckerberg just announced that Threads is taking the next step towards becoming interoperable with the fediverse. Beginning today, people ages 18 and up with public Threads profiles can turn on sharing to the fediverse as part of our beta experience. We're testing this in a few countries to start, including the US.

The fediverse is an open, decentralized social media network that lets people connect and share information across independent servers built on the ActivityPub protocol. If you opt into the fediverse beta experience on Threads, people on other servers will be able to search for and follow your Threads profile, view and interact with your posts, and share your posts with anyone on or off their server. While likes from the fediverse will be aggregated on Threads, you won't be able to see replies or follows from the fediverse in Threads. At this time, polls and posts with reply controls cannot be shared to the fediverse.

If you are a user on another server in the fediverse that communicates with Threads, you can follow Threads users who share to the fediverse and interact with their content. Your server will share information with Threads – for example, if you like posts from Threads users on another server, that information will be shared with Threads.

To turn on sharing to the fediverse, go to your Account Settings, tap on Fediverse sharing (Beta) and follow the instructions. Visit our Help Center and the Threads Supplemental Privacy Policy to learn more about Threads and the fediverse.



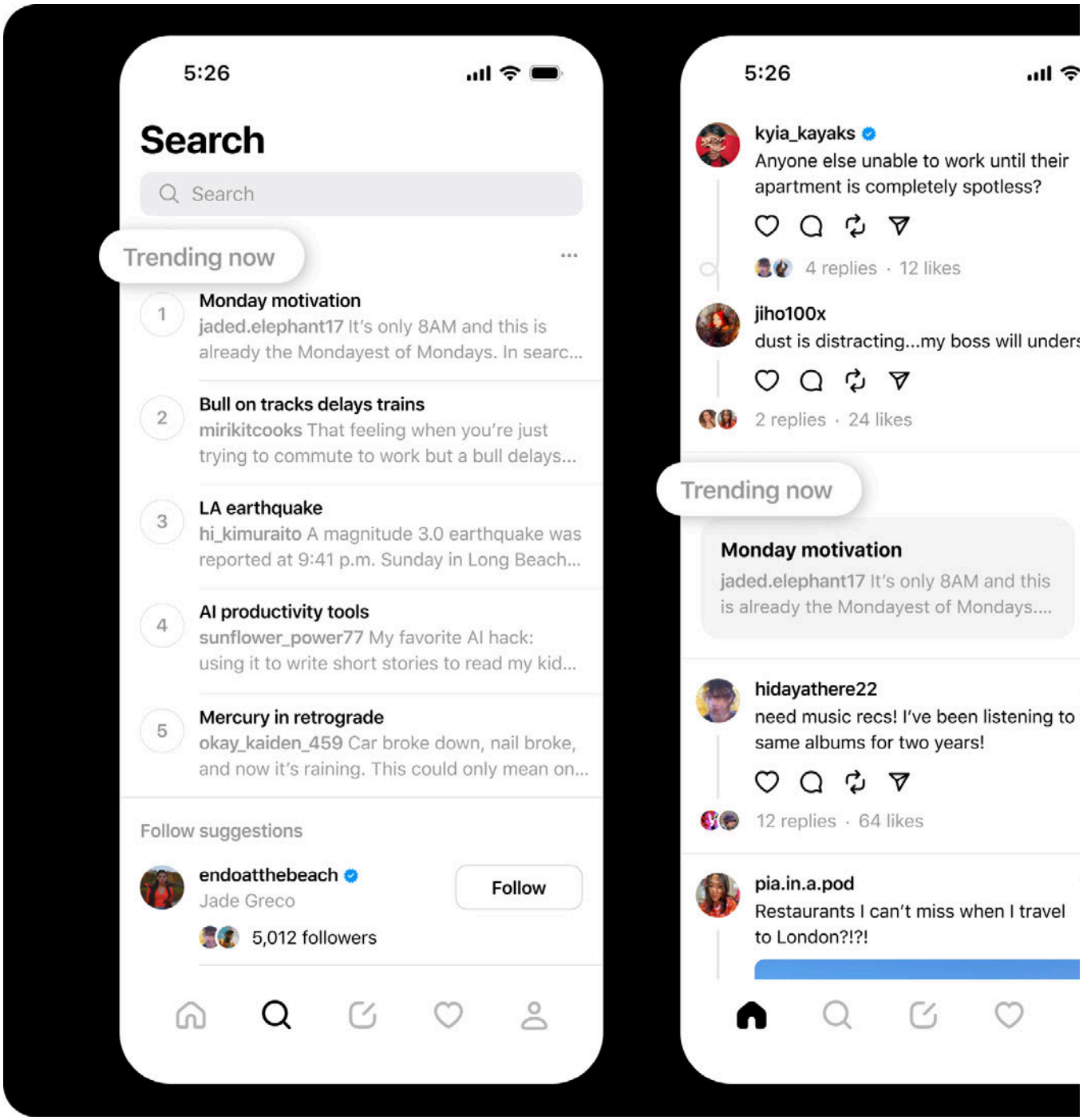


Update on March 19, 2024 at 11:00AM PT:

Adding a Trending Section on Threads

Mark Zuckerberg just announced that we are rolling out a Trending Now feature on Threads, to help you find timely topics and join conversations you care about. At this time, the feature is only available in the US. You can find Trending Now on the search page and in your For You feed on Threads.

Our AI systems identify trending topics based on a number of signals including how many posts are made about a given topic, and how many people have engaged with those posts. After the topics are selected, a team of content specialists review the topics to ensure nothing violates our Community Guidelines or Recommendation Guidelines, and that themes are not duplicative, nonsensical, or misleading. We will show up to five topics that are trending across the US. When you select a topic, you'll be taken to search results for that topic – and can explore and engage in the conversation more easily.



Update on December 14, 2023 at 3:05AM PT:

Today, Mark Zuckerberg announced that we are expanding Threads to countries across Europe. We're excited to give more people the opportunity to follow and join the conversations they care about. Since launching Threads in July, we've made significant improvements to the app, including a web experience, a Following Feed, the ability to edit a post, search with keywords, tag topics and more.

People in the EU can choose to create a Threads profile that is connected to their Instagram account – which means they get the same experience as everyone else around the world – or use Threads without a profile. People who use Threads without a profile can browse content on Threads, search for accounts, share content via link copying or platform sharing, and report Threads content, but can't create a post or interact with content.

We're excited to see more people using Threads and will continue to listen to community feedback to further improve the experience for everyone.

Update on December 12, 2023 at 10:00AM PT:

Fact-Checking on Threads

Early next year, our third-party fact-checking partners will be able to review and rate false content on Threads. Currently, when a fact-checker rates a piece of content as false on Facebook or Instagram, we extend that fact-check rating to near-identical content on Threads, but fact-checkers cannot rate Threads content on its own.

We recently gave Instagram and Facebook users more controls, allowing them to decide how much sensitive or, if they're in the US, how much fact-checked content they see on each app. Consistent with that approach, we're also bringing these controls to Threads to give people in the US the ability to choose whether they want to increase, lower or maintain the default level of demotions on fact-checked content in their Feed. If they choose to see less sensitive content on Instagram, that setting will also be applied on Threads.

Update on October 26, 2023 at 12:00PM PT:

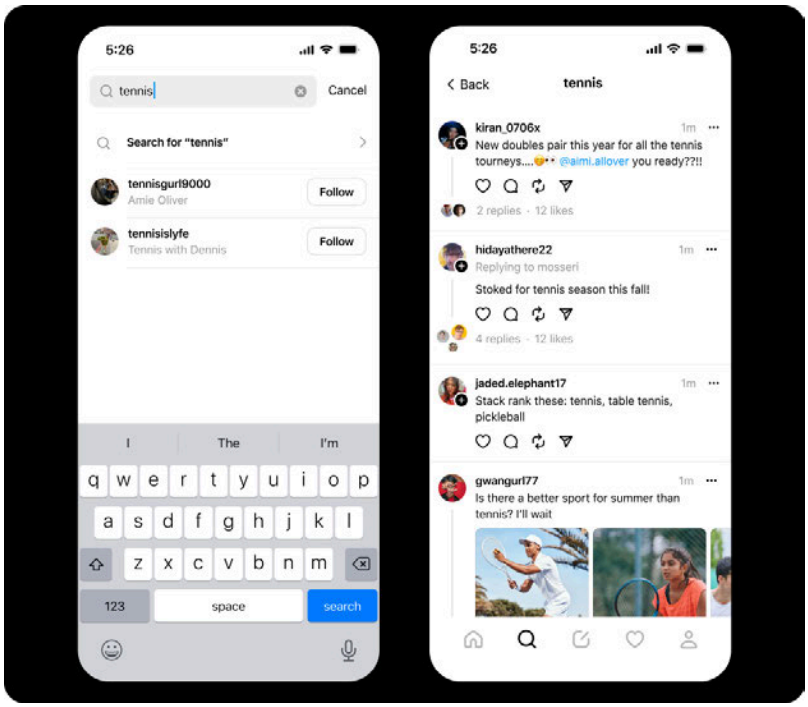
We're adding two new ways to spark conversation on Threads using polls and GIFs.

With polls, you can pose a question with up to four answer choices. The poll will run for 24 hours and you can control who can respond to the poll, the same way you can control who can reply to a post. You have to vote to see current results – and once you vote, you'll be notified when the poll ends to see the final results and join the conversation.

We're also making it easier to share GIFs on Threads. In a new post or reply, select the GIF icon to bring up the GIF library, where you can choose from what's trending or search for a specific GIF.

Update on September 7, 2023 at 11:00AM PT:

Last week, Mark Zuckerberg announced that we began testing keyword search on Threads in Australia and New Zealand. Today, we'll start rolling out keyword search in English and Spanish, in countries where most people post in those languages – such as Argentina, India, Mexico, the UK, and the US – on both mobile and web. We're working on bringing the feature to other languages and countries as soon as we can. And we're continuing to listen to your feedback to improve the search experience.



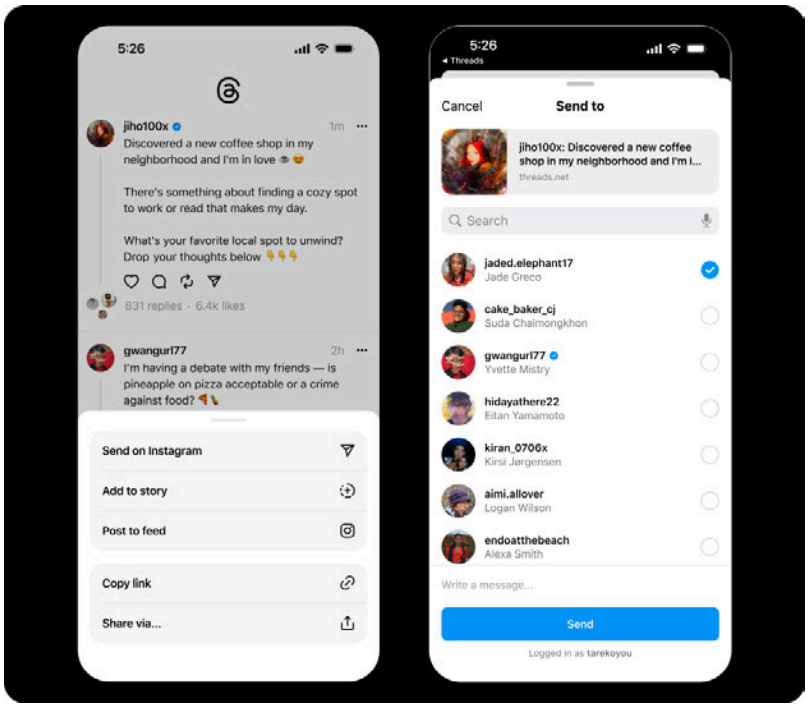
Update on August 22, 2023 at 8:00AM PT:

Mark Zuckerberg just announced that we're rolling out a web experience for Threads over the coming days. The new logged-in experience for web lets you post a Thread, view your feed, and interact with other people's Threads. We're working hard to bring this experience to parity with mobile and will add more functions to the web version in the coming weeks.

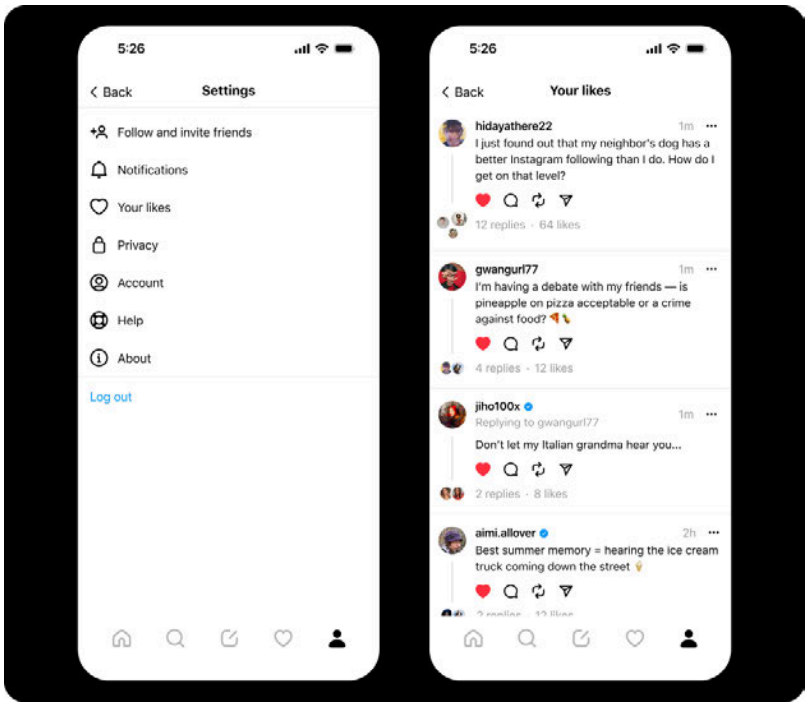
Update on August 9, 2023 at 10:00AM PT:

We've been working on adding new Threads features as quickly as possible, and Mark Zuckerberg just announced that we're rolling out three new ones. You can now:

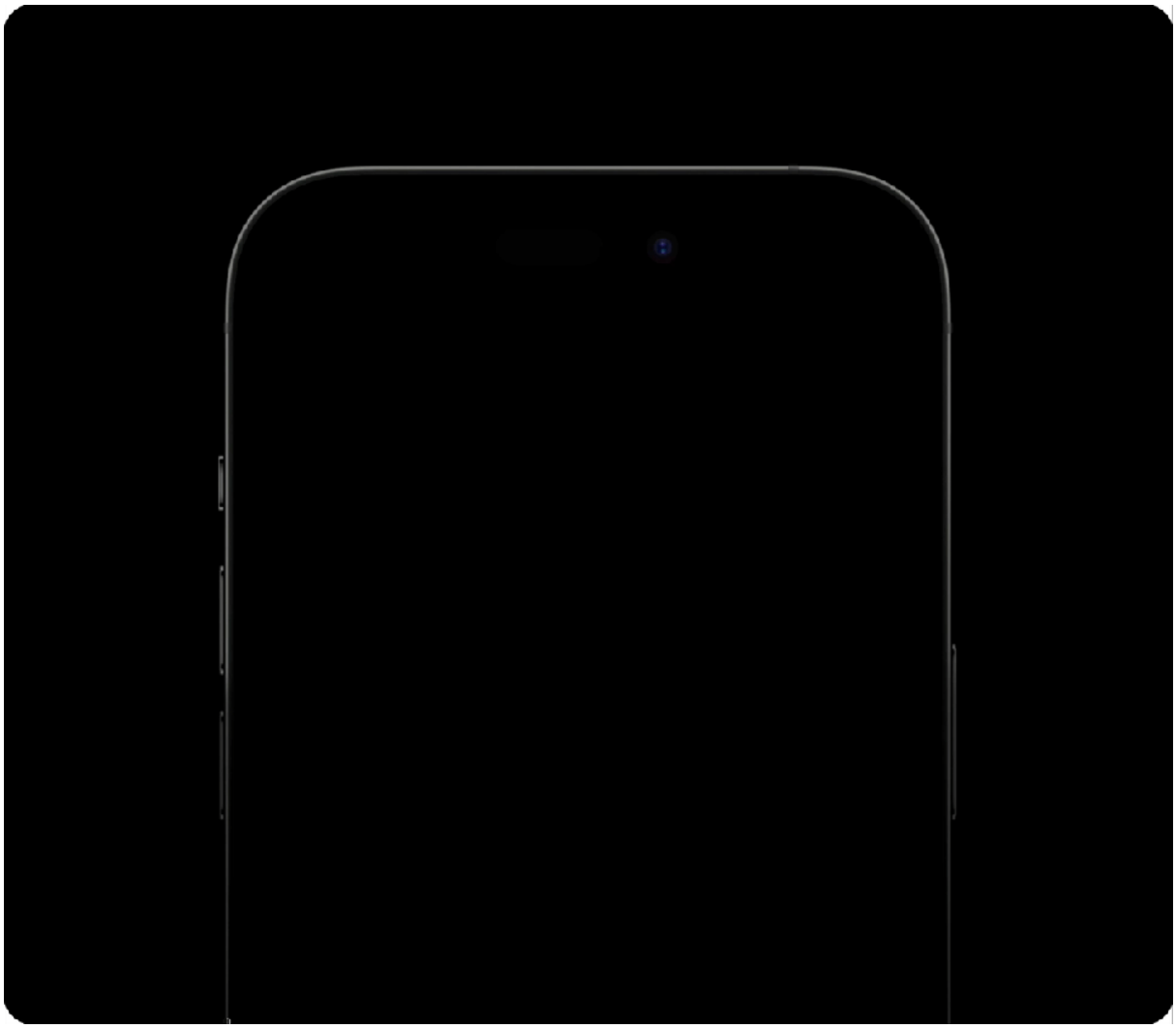
- Share a thread to Instagram DMs
- Use the new mention button to easily mention people in your Thread
- Edit the automatically generated alt-text description of photos and videos before posting



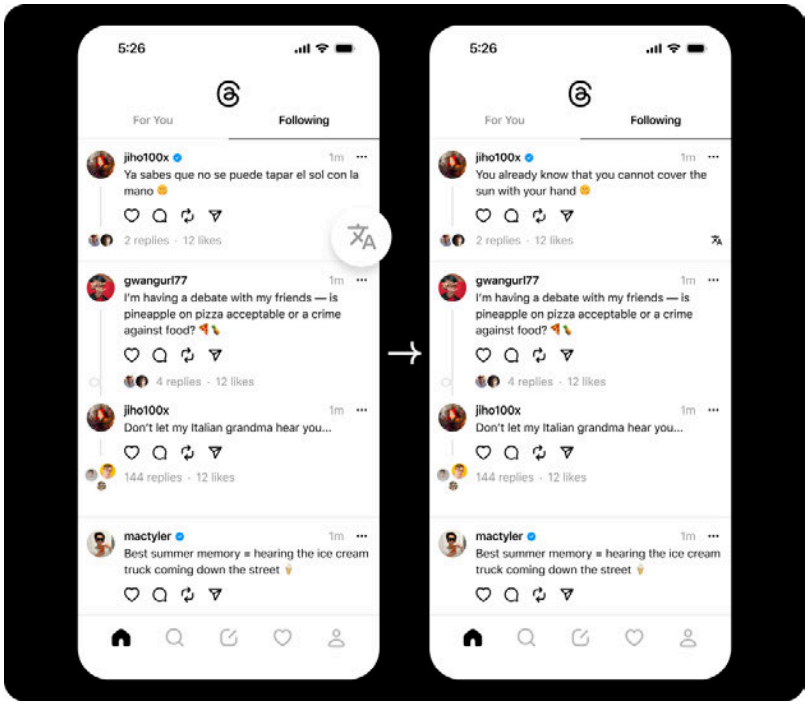
We launched the ability to view posts you've liked in your settings and sort your followers by latest first or earliest first.



We also recently added an option for a chronological feed of only people you're following. So now you can see both: a **For you** feed that includes a mix of posts from profiles you follow and recommended content, as well as a **Following** feed that only shows posts from people you follow.



Threads posts are also now automatically translated based on the language they're written in and the language setting of the person viewing it. If you see a thread in a different language, and your language is available as a translation, you can tap the translation button at the bottom right of the post or reply to see it.



We'll continue listening to feedback and launching more features that improve people's experience on Threads.

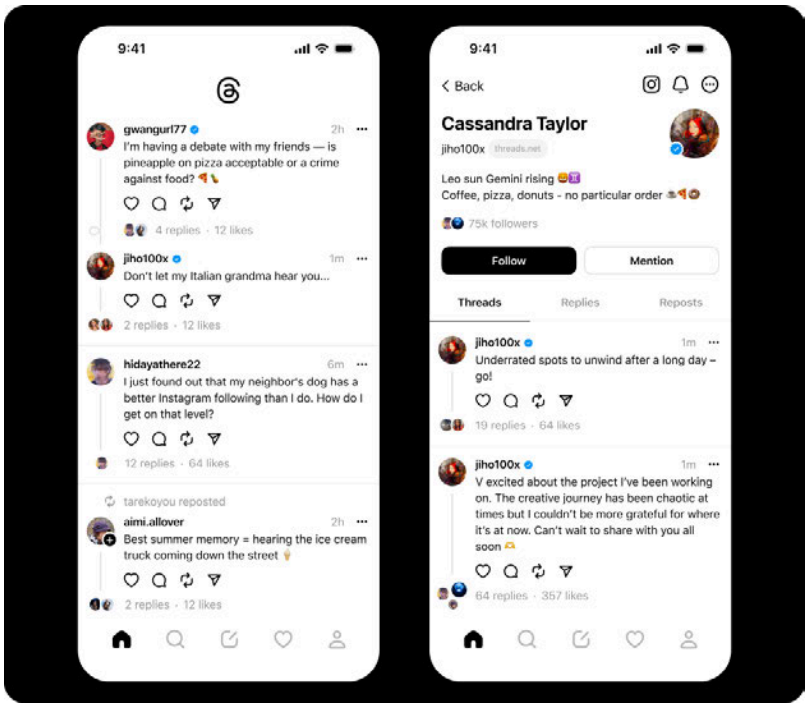
Originally published on July 5, 2023 at 4:30PM PT:

Mark Zuckerberg just announced the initial version of Threads, an app built by the Instagram team for sharing with text. Whether you're a creator or a casual poster, Threads offers a new, separate space for real-time updates and public conversations. We are working toward making Threads compatible with the open, interoperable social networks that we believe can shape the future of the internet.

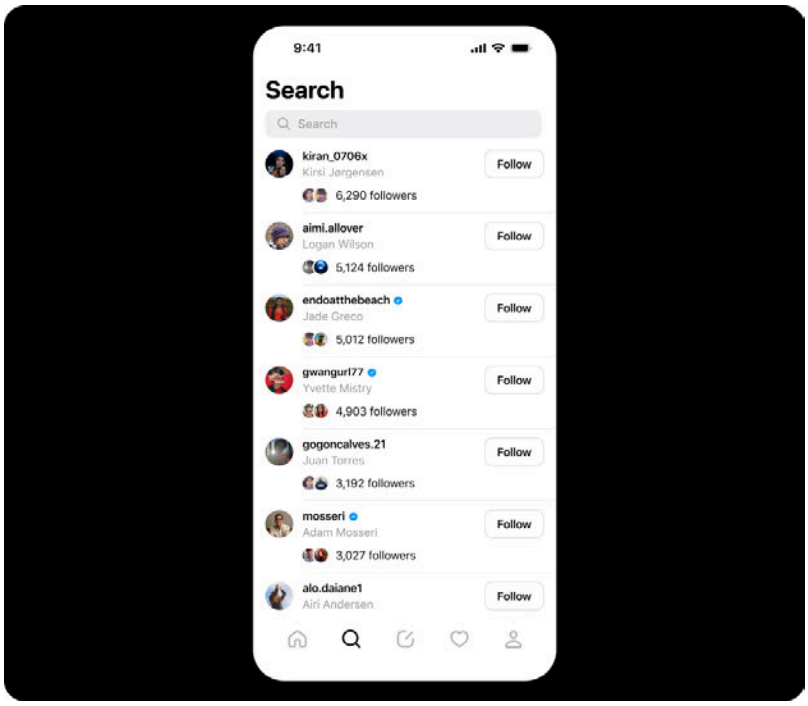
Instagram is where billions of people around the world connect over photos and videos. Our vision with Threads is to take what Instagram does best and expand that to text, creating a positive and creative space to express your ideas. Just like on Instagram, with Threads you can follow and connect with friends and creators who share your interests – including the people you follow on Instagram and beyond. And you can use our existing suite of safety and user controls.

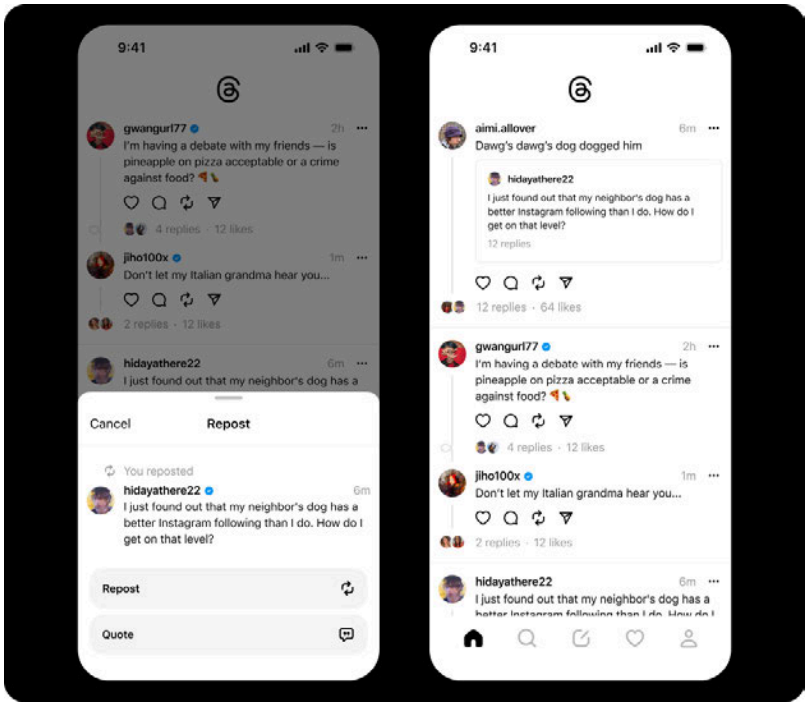
Join the Conversation from Instagram

It's easy to get started with Threads: simply use your Instagram account to log in. Your Instagram username and verification will carry over, with the option to customize your profile specifically for Threads.



Everyone who is under 16 (or under 18 in certain countries) will be defaulted into a private profile when they join Threads. You can choose to follow the same accounts you do on Instagram, and find more people who care about the same things you do. The core accessibility features available on Instagram today, such as screen reader support and AI-generated image descriptions, are also enabled on Threads.





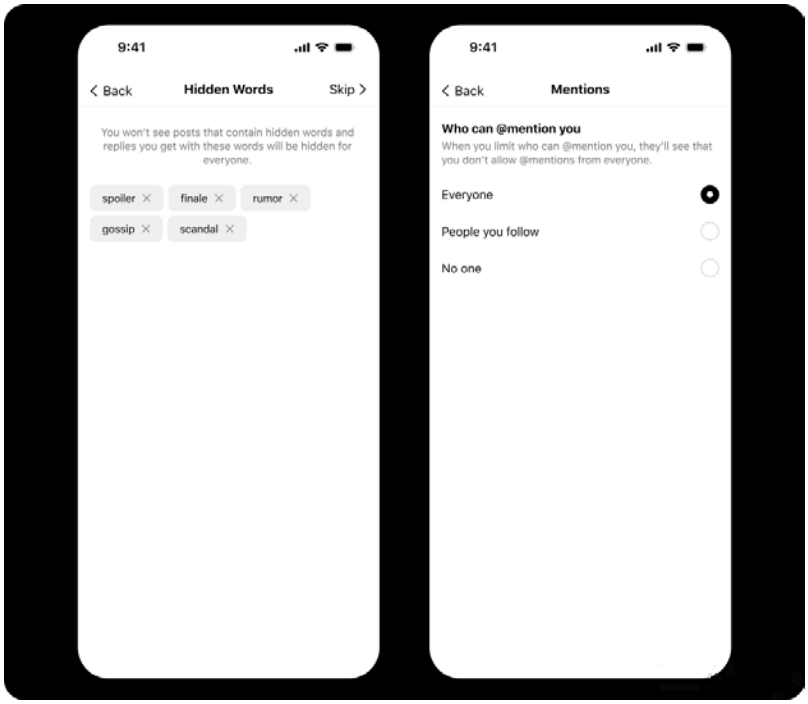
Your feed on Threads includes threads posted by people you follow, and recommended content from new creators you haven't discovered yet. Posts can be up to 500 characters long and include links, photos, and videos up to 5 minutes in length. You can easily share a Threads post to your Instagram story, or share your post as a link on any other platform you choose.

Tune Out the Noise

We built Threads with tools to enable positive, productive conversations. You can control who can mention you or reply to you within Threads. Like on Instagram, you can add hidden words to filter out replies to your threads that contain specific words. You can unfollow, block, restrict or report a profile on Threads by tapping the three-dot menu, and any accounts you've blocked on Instagram will automatically be blocked on Threads.

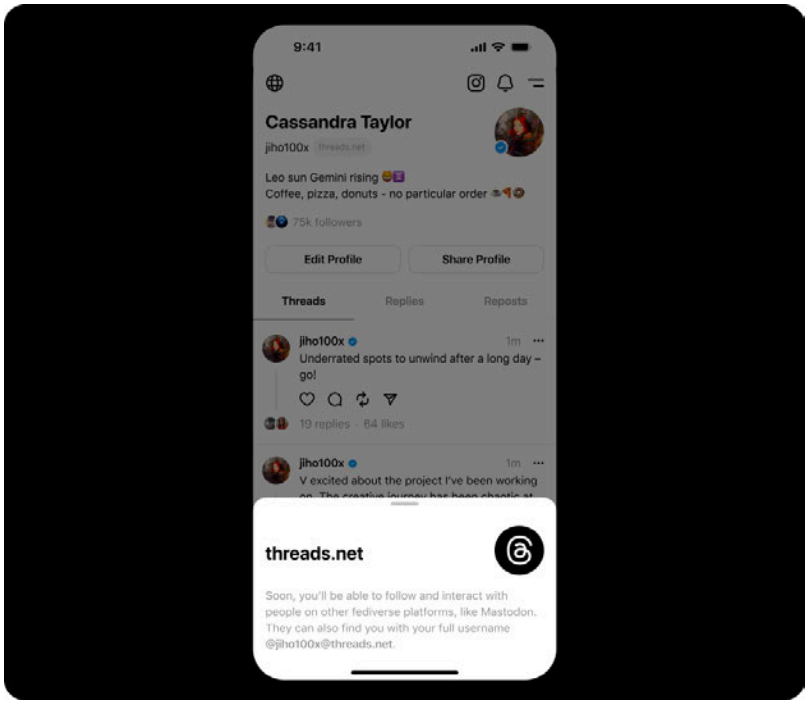
As with all our products, we're taking safety seriously, and we'll enforce Instagram's Community Guidelines on content and interactions in the app. Since 2016 we've invested more than $16 billion in building up the teams and technologies needed to protect our users, and we remain focused on advancing our industry-

leading integrity efforts and investments to protect our community.



Compatible with Interoperable Networks

Soon, we are planning to make Threads compatible with ActivityPub, the open social networking protocol established by the World Wide Web Consortium (W3C), the body responsible for the open standards that power the modern web. This would make Threads interoperable with other apps that also support the ActivityPub protocol, such as Mastodon and WordPress – allowing new types of connections that are simply not possible on most social apps today. Other platforms including Tumblr have shared plans to support the ActivityPub protocol in the future.



We're committed to giving you more control over your audience on Threads – our plan is to work with ActivityPub to provide you the option to stop using Threads and transfer your content to another service. Our vision is that people using compatible apps will be able to follow and interact with people on Threads without having a Threads account, and vice versa, ushering in a new era of diverse and interconnected networks. If you have a public profile on Threads, this means your posts would be accessible from other apps, allowing you to reach new people with no added effort. If you have a private profile, you'd be able to approve users on Threads who want to follow you and interact with your content, similar to your experience on Instagram.

The benefits of open social networking protocols go well beyond the ways people can follow each other. Developers can build new types of features and user experiences that can easily plug into other open social networks, accelerating the pace of innovation and experimentation. Each compatible app can set its own community standards and content moderation policies, meaning people have the freedom to choose spaces that align with their values. We believe this decentralized approach, similar to the protocols governing email and the web itself, will play an important role in the future of online platforms.

Threads is Meta's first app envisioned to be compatible with an open social networking protocol – we hope that by joining this fast-growing ecosystem of interoperable services, Threads will help people find their community, no matter what app they use.

What's Next

We're rolling out Threads today in more than 100 countries for iOS and Android, and people in those countries can download the app from the Apple App Store and Google Play Store.

In addition to working toward making Threads compatible with the ActivityPub protocol, soon we'll be adding a number of new features to help you continue to discover threads and creators you're interested in, including improved recommendations in feed and a more robust search function that makes it easier to follow topics and trends in real time.

We're excited to hear your feedback as we work to build new features and introduce fun new ways to connect on the app.

PX0312

Log in

mikedavismma 07/05/2023

Anyone think this can become bigger than twitter!?

2,819 replies · 1,094 likes

zuck 07/05/2023

It'll take some time, but I think there should be a public conversations app with 1 billion+ people on it. Twitter has had the opportunity to do this but hasn't nailed it. Hopefully we will.

2,081 replies · 11,575 likes

Log in to like, reply and post. Log in

punchup18 07/06/2023

Only if you're moderation team will call users out for behavior AND be able to understand context.

2 likes

maureennjnyc 07/06/2023

For the first time, I think I might actually start liking you. Let's do this!!!!

3 likes

thomas_rosenstand 07/06/2023

Desktop, hashtags and you might get there.

1 like

joanna.alfakih 07/06/2023

PX0312-001

Love the confidence

1 like

_michael_james1 07/13/2023

With no ads.

pussar.io 01/07/2024

yes
<>
| -- ${billboards} -- > §
<//>
*

httuhin.khan 07/22/2023

My account share Please zuck

permi.in 07/10/2023

Did you try for threads.com or at least attempting to have in the future? A .com domain is a differentiator especially if you want the app to be a public conversations app for 1 billion+ people.

@mosseri @gruber

manubazan.ok 07/06/2023

1 like

cesarromeroar 07/06/2023

@zuck falta el botón para traducir.

PX0312-002

Log in

weirdolewsers 07/06/2023
Me cagué de risa

1 like

icebirdstudio 07/06/2023
Can we have a way to search keywords? I hope that's to come...

3 likes

fineboypyl 07/06/2023
Abeg o. My Threads just dey load i no understand o , be like my Instagram too . Wetin dey sup o

1 like

warrior sf 07/07/2023
I'll always remember Mark @zuck wore a T shirt to the opening bell for Facebook, meanwhile everyone else wore suits. I never felt guilty again about wearing T shirts to a special occasion.

PX0312-003

Log in

cynthialynettehailey 07/11/2023

I'm looking forward to the updates and ad policies, I hate ads

temithegreat 07/06/2023

That's fine. We're all for public conversations as long as you don't promote censorship on here like you did on IG and FB. Let's give room for fact-checking and education rather than restricting or suppressing accounts if people question things like "identifying as a cat" or when someone says they don't trust the vaccine. You can be better, Zuck. This can be the new e utopia if you let it.

4 likes

cynthialynettehailey 07/11/2023

I mean it's Meta, same rules apply everywhere I guess

eliiaraujo 07/06/2023

Já deveria ter tradutor

15 likes

jeffhermanson 07/07/2023

Just don't cut off political conversations because they don't generate revenue.

1 like

bleuwoof 07/06/2023

PX0312-004

Case 1:20-cv-03590-JEB Document 347-2 Filed 05/24/24 Page 52 of 512

Log in

1 like

mariaayllon11 07/06/2023

Falta traductor

5 likes

ankurverma01 07/06/2023

There's something special about writing crisp and clear!

6 likes

Log in to see more replies. Log in

PX0312-005

Case 1:20-cv-03590-JEB Document 347-2 Filed 05/24/24 Page 53 of 512

Log in

© 2024 Threads Terms Privacy Policy Consumer Health Privacy Policy Cookies Policy Report a problem

PX0313

This is what Instagram's upcoming Twitter competitor looks like

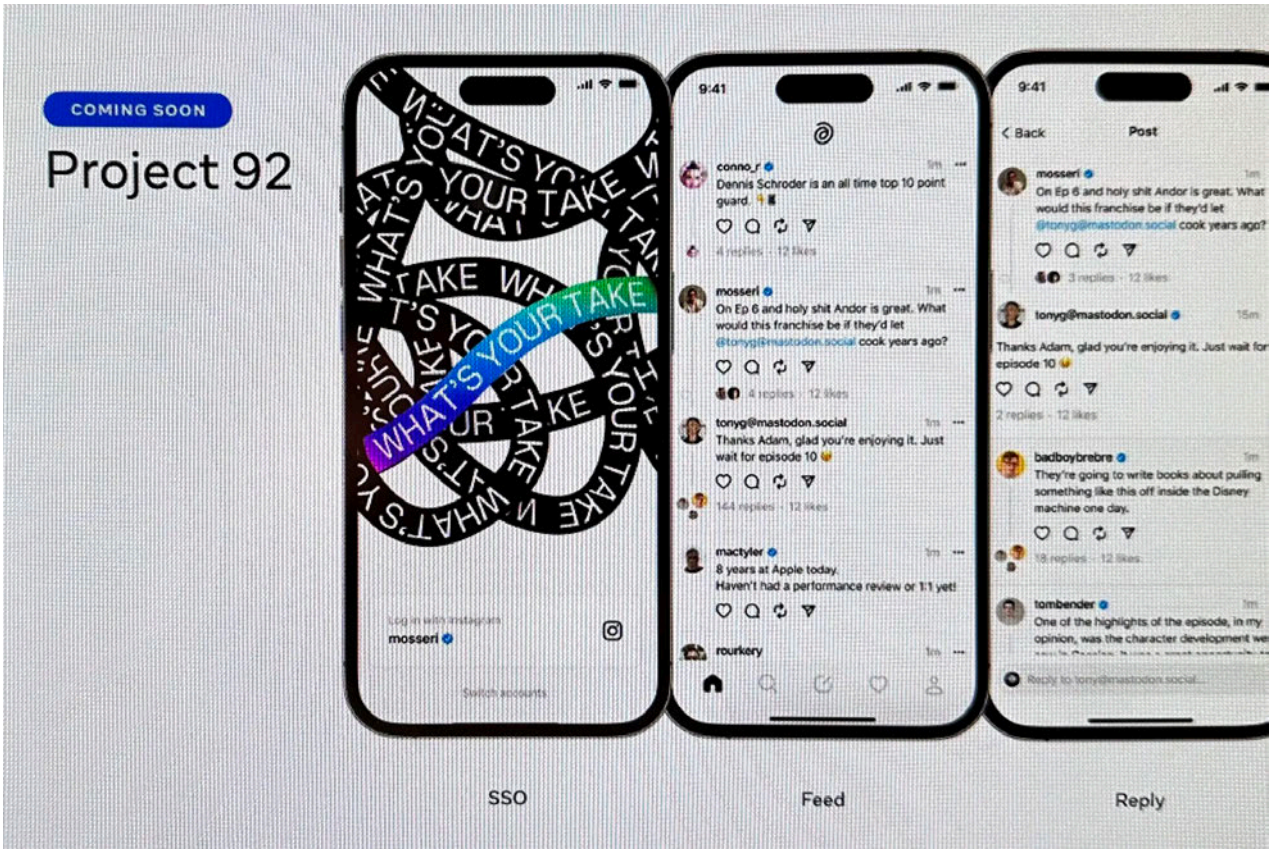
This is what Instagram's upcoming Twitter competitor looks like

/

Meta is talking to celebrities like Oprah and the Dalai Lama about being early users. 'We've been hearing from creators and public figures who are interested in having a platform that is sanely run,' a top exec told employees.

By Alex Heath, a deputy editor and author of the Command Line newsletter. He's covered the tech industry for over a decade at The Information and other outlets.
Jun 8, 2023, 3:51 PM EDT

Share this story



Today, Meta employees were shown these screenshots of Instagram's upcoming app to compete with Twitter.

Image: The Verge

One of Meta's top executives showed employees a preview of the company's upcoming Twitter competitor during a companywide meeting today that was watched by *The Verge*. You can see some of the screenshots above.

The new standalone app will be based on Instagram and integrate with ActivityPub, the decentralized social media protocol. That will theoretically allow users of the new app to take their accounts and followers with them to other apps that support ActivityPub, including Mastodon.

The forthcoming app, which, in the meeting today, Meta chief product officer Chris Cox called "our response to Twitter," will use Instagram's account system to automatically populate a user's information. The internal codename for the app is "Project 92," and its public name could be Threads, based on internal documents also seen by *The Verge*.

"We've been hearing from creators and public figures who are interested in having a platform that is sanely run"

"We've been hearing from creators and public figures who are interested in having a platform that is sanely run, that they believe that they can trust and rely upon for distribution," Cox said, throwing direct shade at Elon Musk's handling of Twitter, to cheers from the audience. He said the company's goal for the app was "safety, ease of use, reliability" and making sure that creators have a "stable place to build and grow their audiences."

Cox said the company already has celebrities committed to using the app, including DJ Slime, and was in discussions with other big names, including Oprah and the Dalai Lama.

He said "coding began" for the app in January and that Meta will be making the app available "as soon as we can."

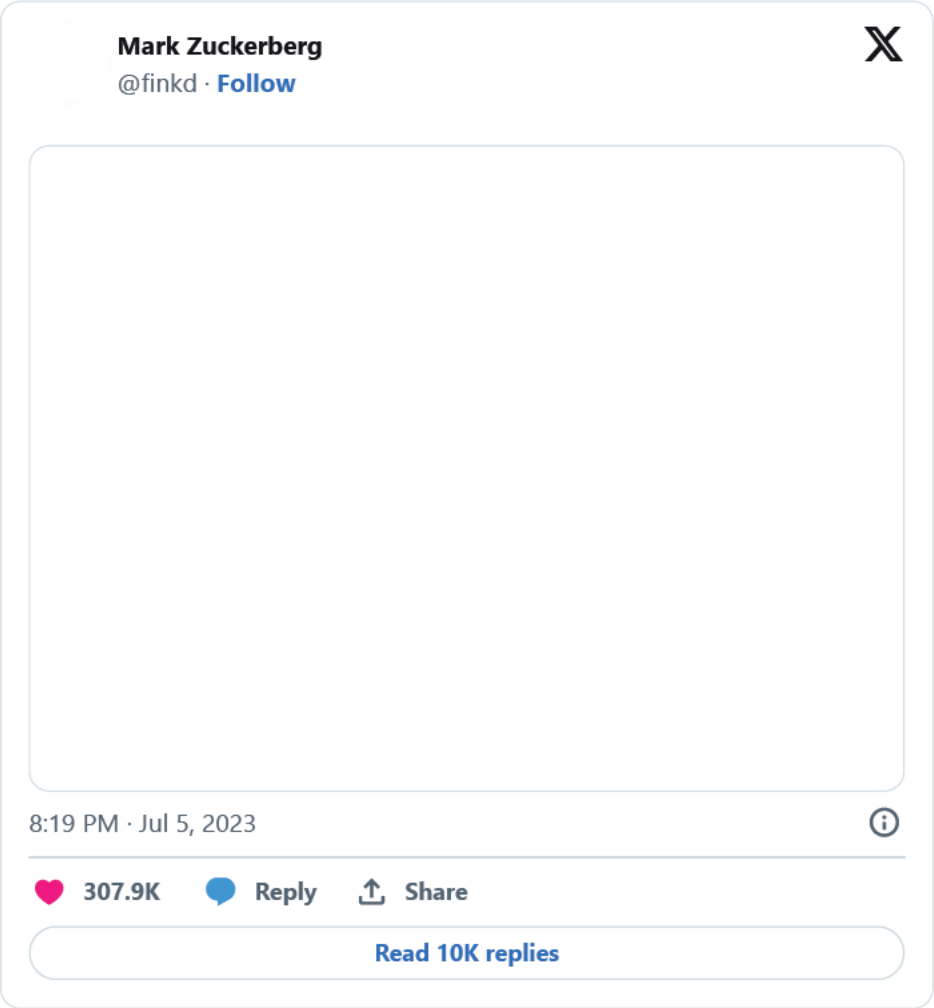
PX0314

Menu

Threads and the Social/Communications Map

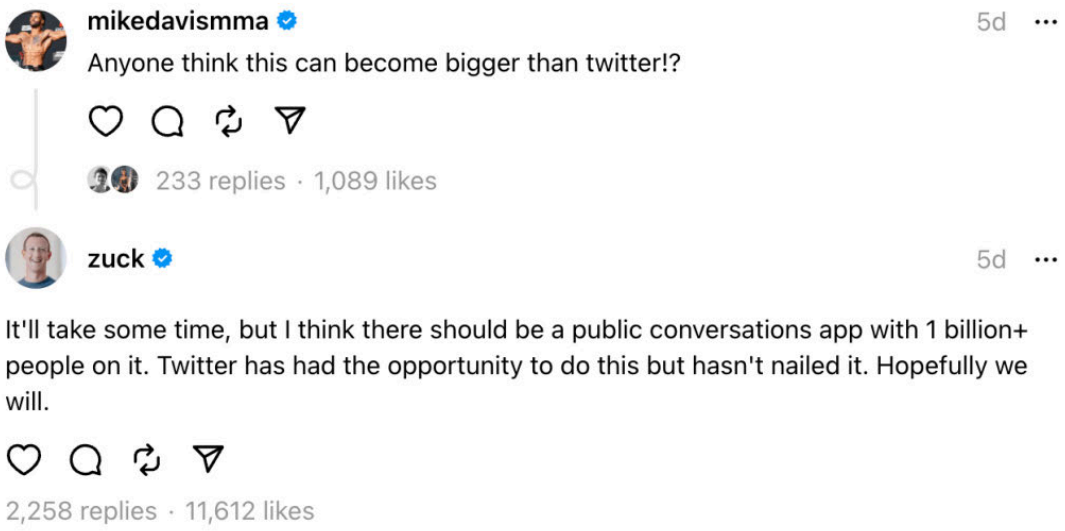
Tuesday, July 11, 2023

If you're only going to tweet once every 11 years, then you better make it count; the best way to do just that is to pull off a well-executed meme:



This offering from Meta CEO Mark Zuckerberg works on multiple levels. The surface interpretation is obvious given the timing of the tweet, which was posted just hours after Meta launched Threads, a text-based social network built on top of the Instagram graph: Threads is a Twitter clone.

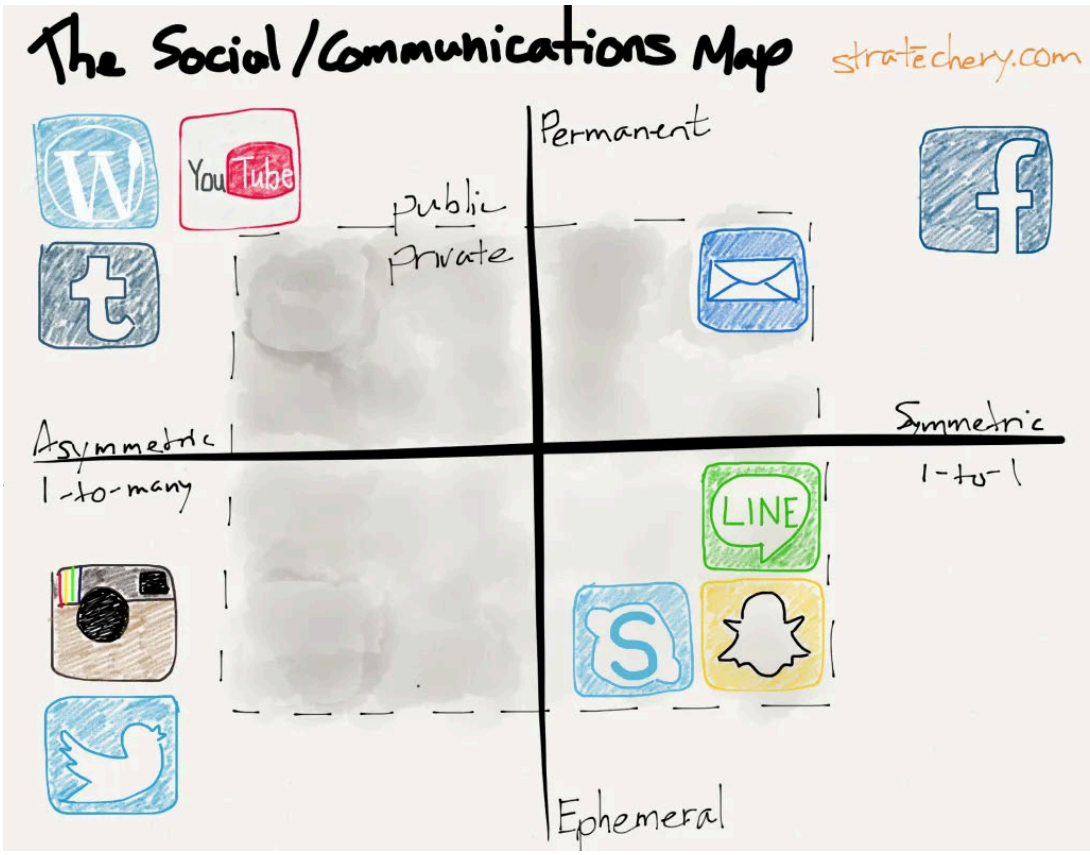
The scene from which the meme is derived, though, gets at what I think is really going on: the Spider-Man on the right is Charles Cameo, an imposter who uses disguises to steal art treasures. To extend the analogy, Threads looks like Twitter at first glance, but is in fact something much different — and what it is stealing is certainly what Elon Musk and Twitter have always wanted:



The important takeaway is that all of the levels of the meme are connected: Threads looks like Twitter, but its essential differences are almost certainly table-stakes in being something larger than Twitter ever was. The question is if that treasure is itself a mirage.

The Social/Communications Map of 2013

Back in 2013 I created The Social/Communications Map:



That map came out of an Article called The Multitudes of Social that argued that social media was not a single category destined to be won by a single app, and that Facebook could never "own social":

> The very idea of owning social is a fool's errand. To be social is to be human, and to be human is, as Whitman wrote, to contain multitudes. Multitudes of apps, in my case...
>
> Facebook needs to appreciate that their dominance of social on the PC was an artifact of the PC's lack of mobility and limited application in day-to-day life. Smartphones are with us literally everywhere, and there is so much more within us than any one social network can capture.

The point about there being a multitude of ways to communicate online has held up well; I think, though, the axis about permanence versus ephemerality was less important than it seemed when the big battle was between Facebook and Snapchat. A better axis leans into the "SocialCommunications" aspect of the title: the most important new social networks of the last few years have been notable for not really being social networks at all.

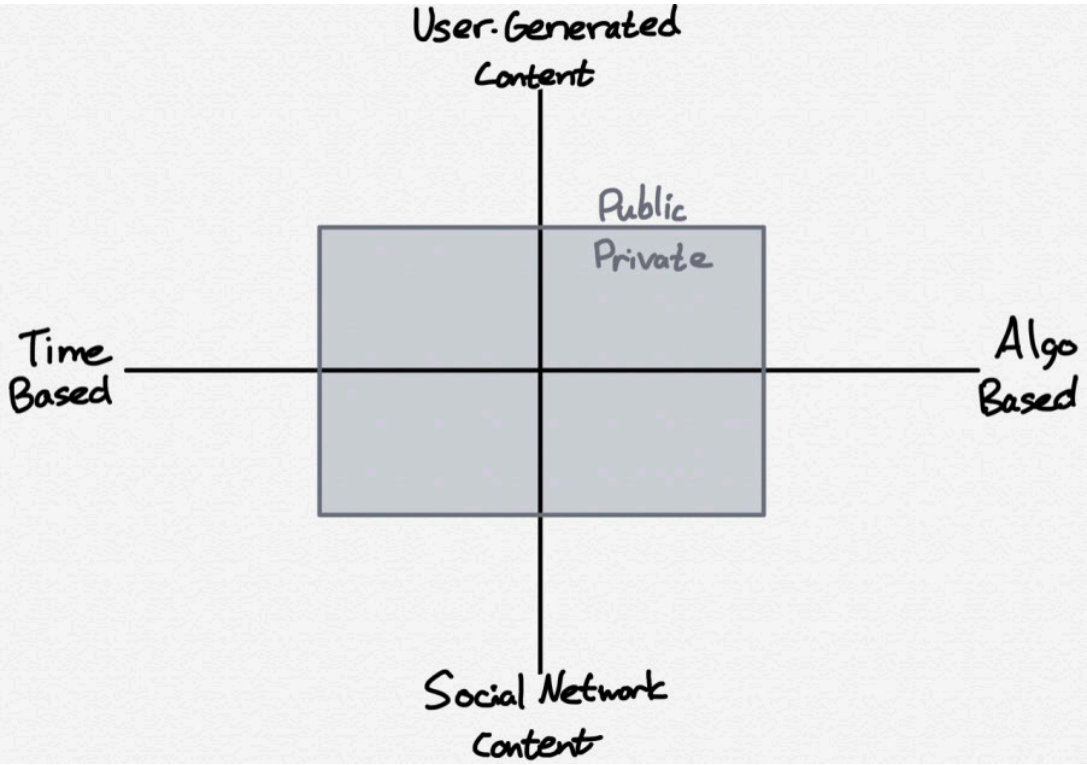
I'm referring to the TikTok-ization of user-generated content: the reason why TikTok was such a blindspot for Facebook is that, unlike Snapchat, it doesn't depend on network effects, but rather abundance. One of the first times I wrote about TikTok was in the context of Quibi, the failed mobile video app from Hollywood impresario Jeffrey Katzenberg:

> The single most important fact about both movies and television is that they were defined by scarcity: there were only so many movies that would ever be made to fill only so many theater slots, and in the case of TV, there were only 24 hours in a day. That meant that there was significant value in being someone who could figure out what was going to be a hit before it was ever created, and then investing to make it so. That sort of selection and production is what Katzenberg and the rest of Hollywood have been doing for decades, and it's understandable that Katzenberg thought he could apply the same formula to mobile.
>
> Mobile, though, is defined by the Internet, which is to say it is defined by abundance...So it is on TikTok, or any other app with user-generated content. The goal is not to pick out the hits, but rather to attract as much content as possible, and then algorithmically boost whatever turns out to be good...The truth is that Katzenberg got a lot right: YouTube did have a vulnerability in terms of video content on mobile, in part because it was a product built for the desktop; TikTok, like Quibi, is unequivocally a mobile application. Unlike Quibi, though, it is also an entertainment entity predicated on Internet assumptions about abundance, not Hollywood assumptions about scarcity.

It's ultimately a math question: are you more likely to find compelling content from the few hundred people in your social network, or from the millions of people posting on the service? The answer is obviously the latter, but that answer is only achievable if you have the means of discovering that compelling content, and, to be fair to both Facebook and Twitter, the sort of computational power necessary to pull off a TikTok-style network didn't exist when those companies got started.

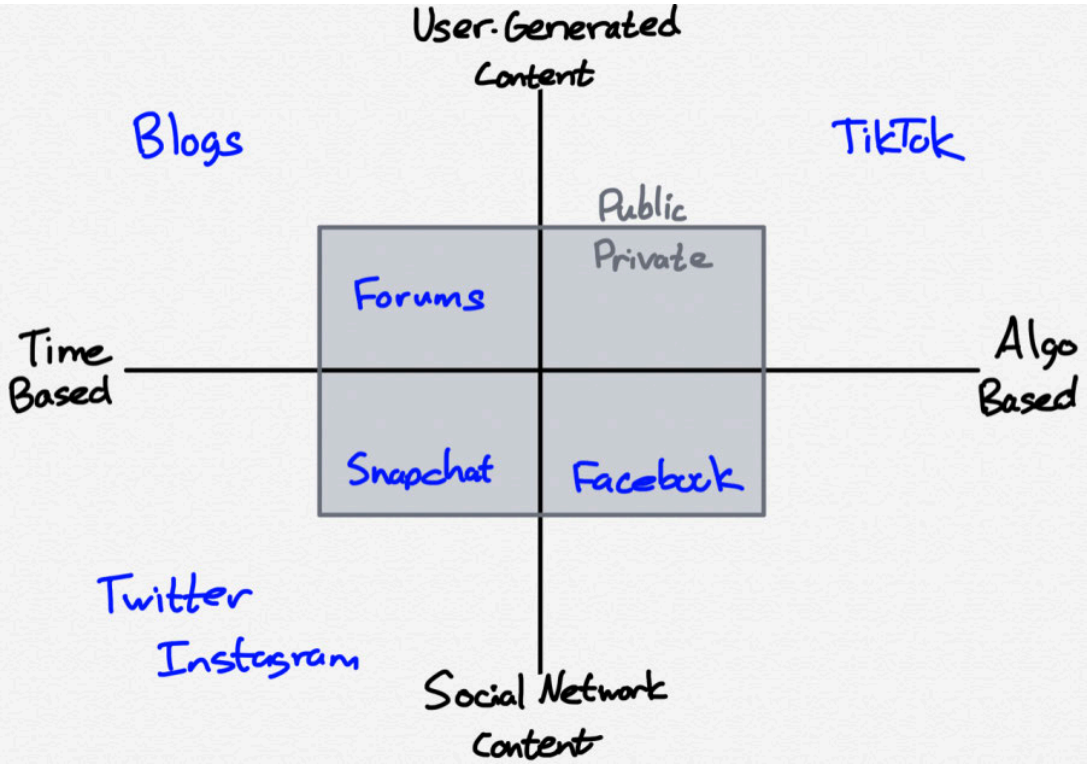
The Social/Communications Map of 2023

Set that point about time of origin aside just for a moment; here is what I think a better representation of the Social/Communications Map looks like in 2023:



The first change is that the symmetric/asymmetric axis has been replaced by the nature of the sorting algorithm: chronological order versus algorithmic selection. However, this isn't that big of a change; consider messaging, which is by definition about symmetric social networking. Messaging only really makes sense if it is organized by time — imagine trying to carry on a conversation if every message you saw were algorithmically selected, instead of simply displayed in order. Algorithmic sorting, though, makes much more sense when you are consuming content that is broadcast to the world, and thus has no assumptions about or expectations for in-order contextual replies.

The second change is the TikTok-ization I noted above: my new vertical axis is user-generated content, by which I mean content across the network, versus network-generated content, by which I mean content from the people you choose to follow. If you maintain the same public/private distinction I had in the original, you get a landscape that looks something like this (note that Facebook is better thought of as a private social network, given that the default nature of posts is that they are only seen by those in your network).



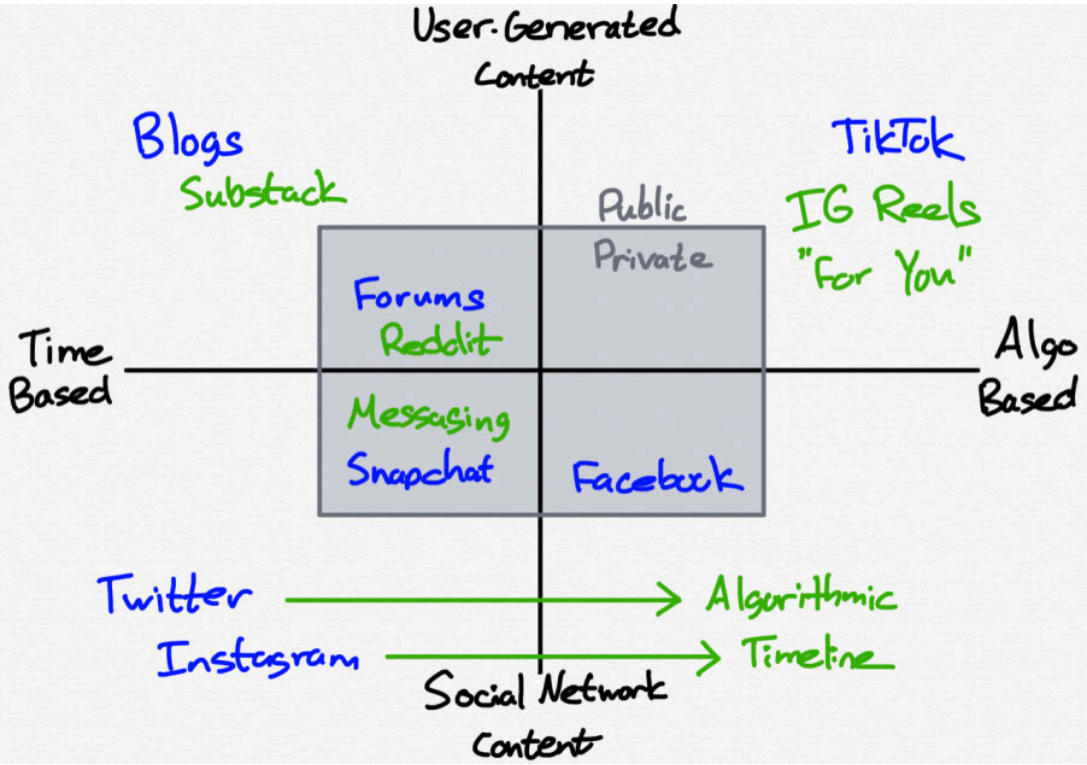
This is where the bit above about historical time comes in: another way to look at this map is as a representation of how content on the Internet has evolved; the early web, and early forms of user-generated content like forums and blogs, were and are still located in the upper left. This quadrant is fairly decentralized, and is Aggregated by Google and search.

The lower left quadrant came next: one site held all of the content from your network, and presented it chronologically. Some sites, like Twitter and Instagram, stayed here for years; Facebook, though, quickly jumped ahead to the lower right quadrant, and organized your feed algorithmically. This quadrant became the other major pillar of Internet advertising (along with search): figuring out what content to show you from your network wasn't too dissimilar of a problem from figuring out what ads to show you, and the nature of a dynamically-generated feed that was unique to every individual was something that was only possible with digital media.

The final stage is, as noted, represented by TikTok: once again your network doesn't matter, because the content comes from anywhere. This world, though, unlike the open web, is governed by the algorithm, not time or search.

Twitter, Threads, and the Upper-Right

I was honestly surprised to find out that both Twitter and Instagram were in the lower left quadrant until 2016; that is when both services started offering an algorithmic timeline. Of course the surprise for the two services ran in the opposite direction: for Twitter it's amazing that the company managed to change anything at all, and for Instagram it's a surprise the service stayed the same for so long. Since then Instagram has heavily invested in its direct messaging product even as it has slowly abandoned the public parts of the lower left: everything is an algorithm and, with Reels, completely disconnected from your network.



Perhaps the starkest change that Musk has made to Twitter, meanwhile, has been a headlong rush into the upper right: the "For You" tab is far more aggressive about promoting tweets from people you don't follow, and it's increasingly impossible to escape; the app always defaults to "For You", and there are no more 3rd-party app alternatives. Eugene Wei argues this has blown up the timeline and ruined the Twitter experience:

> What established the boundaries of Twitter? Two things primarily. The topology of its graph, and the timeline algorithm. The two are so entwined you could consider them to be a single item. The algorithm determines how the nodes of that graph interact. In a literal sense, Twitter has always just been whose tweets show up in your timeline and in what order.
>
> In the modern world, machine learning algorithms that mediate who interacts with whom and how in social media feeds are, in essence, social institutions. When you change those algorithms you might as well be reconfiguring a city around a user while they sleep. And so, if you were to take control of such a community, with years of information accumulated inside its black box of an algorithm, the one thing you might recommend is not punching a hole in the side of that black box and inserting a grenade. So of course that seems to have been what the new management team did. By pushing everyone towards paid subscriptions and kneecapping distribution for accounts who don't pay, by switching a TikTok style algorithm, new Twitter has redrawn the once stable "borders" of Twitter's communities.
>
> This new pay-to-play scheme may not have altered the lattice of the Twitter graph, but it has changed how the graph is interpreted. There's little difference. My For You feed shows me less from people I follow, so my effective Twitter graph is diverging further and further from my literal graph. Each of us sits at the center of our Twitter graph like a spider in its web built out of follows and likes, with some empty space made of blocks and mutes. We can sense when the algorithm changes. Something changed. The web feels deadened.
>
> I've never cared much about the presence or not of a blue check by a user's name, but I do notice when tweets from people I follow make up a smaller and smaller percentage of my feed. It's as if neighbors of years moved out from my block overnight, replaced by strangers who all came knocking on my front door carrying not a casserole but a tweetstorm about how to tune my ChatGPT and MidJourney prompts.

Instagram's Evolution has shown that this shift is possible, but the shift has been systemic and gradual — and even then subject to occasionally intense pushback. Musk's Twitter, though, has been haphazard and blistering in its pace. What ought to concern the company about Threads, though, is the possibility that all of the upheaval — which effectively sacrifices the niche Twitter had carved out amongst text nerds that dominate industries like media — will not actually result in the user growth Musk is hoping for, because Threads got there first.

Indeed, this map is the key to understanding why it is that Threads looks like Twitter, but is in fact a very different product: Threads is solidly planted in the upper right. When you log onto the app for the first time, your feed is populated by the algorithm; there is some context given by whom you follow on Instagram, but Meta seems aware that accounts you might want to look at may be different than accounts you want to hear from, and is thus filling the feeds with what it thinks you might find interesting. That is how it can provide an at-least-somewhat-compelling first-run experience to 100 million people in five days.

Twitter, on the other hand, faces the burden of millions having tried the service in past iterations and quickly deciding it wasn't for them; even if the algorithm were effective, it may already be too late to gain new users, even as you sacrifice what the service's existing users preferred.

The Threads Experiment

This leads to the biggest open question about Threads' long-term prospects, and, by extension, Twitter's: did those millions of abandoned Twitter users give up because text-based social networking just wasn't that interesting to them, or because Twitter made it too hard to get started? I've made the case that it's the former, which means that Threads is a grand experiment as to the validity of that thesis. If those 100 million users stay engaged (and if that number continues to grow), then the people chalking up Twitter's inability to grow or monetize effectively to the company's inability to execute are correct.

At the same time, as Wei notes, Musk's tenure has highlighted the problems with doing too much: what if Twitter succeeded to the extent it did not despite management's seeming ineffectiveness, but because of it?

> I've written before in Status as a Service or The Network's the Thing about how Twitter hit upon some narrow product-market fit despite itself. It has never seemed to understand why it worked for some people or what it wanted to be, and how those two were related, if at all. But in a twist of fate that is often more of a factor in finding product-market fit than most like to admit, Twitter's indecisiveness protected it from itself. Social alchemy at some scale can be a mysterious thing. When you're uncertain which knot is securing your body to the face of a mountain, it's best not to start undoing any of them willy-nilly. Especially if, as I think was the case for Twitter, the knots were tied by someone else (in this case, the users of Twitter themselves).

Many of those knots are tied to that lower left quadrant: a predominantly time-based feed makes sense if a service is predominantly about "What is happening?", to use Twitter's long-time prompt; a graph based on who you choose to follow doesn't just show what you want to see, it also controls what you don't (Wei notes that this is a particularly hard problem for algorithmically generated feeds). Both qualities seem particularly pertinent for a medium (text) that is information dense and favored by people interested in harvesting information, a very different goal than looking to pass the time with an entertaining video or ten.

It follows, then, that Twitter's best defense against Threads may be to retreat to that lower left corner: focus on what is happening now, from people you chose to follow. The problem, though, is that while this might win the battle against

Threads, it means that Musk will have lost the war when it comes to ever making a return on his $44 billion. In truth, though, that war is already lost: Musk's lurch for the upper right was probably the best path to reigniting user growth, but if that is the corner that matters then Threads will win.

Thread's Chronological Timeline

The other question is if Threads will come for Twitter's place on the map; Head of Instagram Adam Mosseri says that a chronological timeline is coming:

Placing this option in the context of Facebook and Instagram actually suggests that this feature won't matter very much; both services make it hard to find, and revert back to the default algorithmic feed, and for good reason: users may say they want a chronological feed, but their revealed preference is the opposite. Instagram founders Kevin Systrom and Mike Krieger, who initially opposed algorithmic ranking in Instagram, told me in a Stratechery Interview:

> Kevin Systrom: I remember thinking when the team was like, "We're thinking of using machine learning to sort the Explore page," I'm not even sure what they call it now, but basically the Explore page and I remember saying, "It just feels like that's a bunch of hocus-pocus that won't work. Or maybe it'll work but you won't really understand what it's doing and you won't fully understand the implications of it, so we should probably just keep it very simple." I was so wrong and I only remember it because I was so wrong, but you asked about feed, Mike would probably give you his anecdote about feed. But on the Explore page I was very anti and then I think I became pro only once I saw what it could do. Not in terms of just usage metrics, but just the quality of what people were served compared to some of our heuristics before...

> **Mike Krieger:** I'll share a funny anecdote about the Explore experiment. Facebook has all these internal A/B testing tooling and we hooked into it and we ran our first machine learning on the Explore experiment and we filed a bug report and I'm like, "Hey, your tool isn't working, that's not reporting results here." And they said, "No, the results are just so strong that they're literally off the charts. The little bars that show it literally is over 200%, you just should ship this yesterday." The data looked really good.

That noted, observe Mosseri's stated goal for the app, as articulated to Alex Heath of The Verge:

> I think success will be creating a vibrant community, particularly of creators, because I do think this sort of public space is really, even more than most other types of social networks, a place where a small number of people produce most of the content that most everyone consumes. So I think it's really about creators more than it is about average folks who I think are much more there just to be entertained. I think [we want] a vibrant community of creators that's really culturally relevant. It would be great if it gets really, really big, but I'm actually more interested in if it becomes culturally relevant and if it gets hundreds of millions of users. But we'll see how it goes over the next couple of months or probably a couple of years.

"Culturally relevant" is the one game that Twitter has won, far more than Facebook, and arguably more than Instagram: Twitter drives national and international media coverage, from TV to newspapers, to an extent that drastically exceeds its monetization potential. Meta, meanwhile, has been content to provide social networking for the silent majority, making tons of money along the way. The best way to do that with text — if it is even possible — would be to stay in that upper right corner; cultural relevancy, though, is still in the bottom left, even if there aren't nearly as many users, or money.

And, it must be noted, Twitter is vulnerable in its home territory; I've long argued that the importance of convenience in terms of app success is underrated (see Threads starting with your Instagram sign-in and network), but its hard to think of anything that might motivate users to make a change more than resolving cognitive dissonance. There is a sizable segment of that culturally relevant audience Mosseri wants to capture who are opposed to Musk, and yet can't give up Twitter; I suspect that much of the outpouring of glee over Threads' early success is from this cohort that wants nothing more than non-Musk Twitter.

Ultimately, though, I think they may be disappointed: Meta is about algorithms and scale, and I would bet that Threads will leave real time reactions, news, and pitched battles to Twitter; Musk's most important decision may be accepting that that is enough, because it's all he's going to get.

← Amazon, Friction, and the FTC
Hollywood on Strike →

Subscriber's Daily Update

Tuesday, April 2, 2024
The XZ Backdoor, What Happened, Open Source Safety
Monday, April 1, 2024
MLS on Vision Pro, The Vision Pro's Missing Content, The Vision Pro's DRI
Thursday, March 28, 2024
An Interview with Hugo Barra About a Career in Tech and the Future of VR
Wednesday, March 27, 2024
Apple and Antitrust for Aggregators, DMA Investigations and Apple's Risk

On the business, strategy, and impact of technology.

© Stratechery LLC 2024 | Terms of Service | Privacy Policy

PX0315

Los Angeles Times

TECHNOLOGY AND THE INTERNET

Threads, Meta's Twitter clone, just launched — and people are already complaining about it

Facebook parent Meta launched a new app that appears to mimic Twitter, marking a direct challenge to the social media platform owned by billionaire Elon Musk.

By Helen Li
Staff Writer

July 6, 2023 **Updated** 2:44 PM PT

Mastodon, Spill, BlueSky — the list of Twitter alternatives continues to grow as Elon Musk's chaotic reign drives users to explore their options. Now comes the big one, from the company with the resources and scale to make a real go of it.

On Wednesday, Meta, the parent company of Facebook and Instagram, launched its Twitter rival app: Threads. Scrolling through the blocks of black on white text, it looks a lot like its rival. And setup is a snap for anyone who already has an Instagram account: Users can create and autopopulate a profile with the same log-in and bring their followers over.

Just like Twitter, the platform is aimed at creating public conversations, according to the company.

"We're hoping to bring some of what we have built for photos and videos on Instagram to Threads with text," said Adam Mosseri, the head of Instagram, in a video. He said content creators already on Instagram would be able to have more "friendly" and "open" conversations with their followers than they would on other platforms — i.e., on Twitter, where pile-ons and other toxic behavior are a frequent complaint.

In a Thread post, Meta Chief Executive Mark Zuckerberg said that Threads gained 10 million users in seven hours and by Thursday morning had 30 million.

TECHNOLOGY AND THE INTERNET

Twitter threatens legal action over Threads, calling Meta's new app a 'copycat'

July 6, 2023

Twitter responded by sending a threatening legal letter to Zuckerberg accusing him of hiring ex-Twitter employees and using their confidential knowledge to build a "copycat" product, according to the news outlet Semafor.

"Competition is fine, cheating is not," Musk tweeted.

The user experience on Twitter has changed dramatically in the months since Musk took over in a $44-billion acquisition and laid off more than half of its employees. Among the major changes has been the introduction of a paid subscription model, Twitter Blue,

whose users pay $8 for perks including more visibility in other users' feeds. Verified users who declined to pay lost their blue check marks, although the company subsequently restored them for celebrities and others with more than 1 million followers.

During the long Fourth of July holiday weekend, users suddenly ran into a limit in the number of tweets they could view as the platform shifted off of Google's cloud storage. Twitter explained the throttling as a necessary measure to "reduce downtime and error pages."

For users hoping that Threads would offer the same experience minus the headaches, the debut has been a mixed picture. Complaints about the new service are already cropping up on Threads itself and other social media platforms.

ADVERTISEMENT

First, there is currently no desktop version, a must for those who want to scroll at work or get off their phones.

The composition of the feed is another sore point, especially for Twitter users repelled by its algorithmic "For You" tab full of clickbait and other viral flotsam. Threads offers no option to view posts only from accounts you follow or to view the feed chronologically.

TECHNOLOGY AND THE INTERNET

Elon Musk is dying to be funny. These eight comedians have some notes

May 24, 2023

Users also have commented on the app's failure to address concerns for users with disabilities.

"Threads was shipped without basic accessibility functions like an alt text field or an in-app captioning tool," Alexa Heinrich said on Twitter.

The search bar does not help much with finding the public conversations Mosseri touts. As of now, users can type and search only account usernames, rather than the content of posts. Hashtags and authors who repost are not clickable.

Mosseri said Threads will soon bring features focused on recommendations and trends. It also eventually will integrate ActivityPub, a decentralized social network that will allow creators to take their followers off-platform.

"If you're a creator, you should own your audience," Mosseri said. "We're working on that and a number of other features to improve Threads as quickly as we can."

Users who try Threads and find it wanting may find untangling themselves more complicated than getting started was. By Thursday morning, many on social media were warning that, although deactivation is easy enough, there's no way to delete the new app completely without also deleting the associated Instagram profile.

MORE TO READ

Facebook, Instagram, Messenger and Threads logins restored after widespread outage

March 5, 2024

PX0316

WHAT'S NEW

What we love and hate about Threads, Meta's new Twitter clone

Threads may be the first Twitter alternative that really matters because it's built on top of Instagram's existing base of billions of users

Analysis by Geoffrey A. Fowler and Naomi Nix

Updated July 12, 2023 at 12:58 p.m. EDT | Published July 5, 2023 at 7:00 p.m. EDT

Mark Zuckerberg has unveiled Threads, a clone of Twitter designed to lure people turned off by the social network's changes under owner Elon Musk.

And in its first five days, 100 million people signed up for the free app. The billionaire social media smackdown is about to get real.

What does Threads mean for you — and should you join in the rain on Musk's parade?

On the one hand, Threads has a decent chance of becoming a — maybe even the — major new hub for text-based online conversations. Unlike other would-be Twitter rivals Bluesky and Mastodon, Threads arrives with a potential audience of billions who already use Meta's photo and video-oriented Instagram, which Threads is built on top of. Meta says it's taking moderation seriously to make Threads a safer place for us (and eventually advertisers). Zuckerberg, too, is less inclined than Musk to put his foot in his mouth.

But Threads also comes with a whole host of Meta baggage, including questionable privacy practices, opaque algorithms and Big Tech monopoly power. Many of those issues have turned people off Zuckerberg's other social networks such as Facebook. And those issues are all still present in Threads. For example, from the moment you first log in to Threads, it starts showing you recommended posts from accounts and brands you don't necessarily follow — or necessarily even care to see.

What's to like — and dislike — in Threads? And how do you give it a try? Answers below. , . Send us an email about what else you would like to know.

Getting on Threads is simple

You'll need to have an Instagram account to sign up for Threads. Then you can download the Threads app for iOS or Android to set up your account. You'll use the same name for your Threads account as you do on Instagram.

When you first use the app, you'll be given the option to automatically follow all of the same accounts you follow on Instagram — or just select some of them. We're curious to see how this plays out: How much overlap is there between accounts you follow on Instagram for their epic photographs with accounts you want to read for their text or hot takes on politics and TV?

That system also means your existing Instagram friends and followers don't automatically follow you on Threads. You'll have to build up that audience all over again.

And one other thing to note: At launch, Threads isn't available in Europe, where the Irish Data Protection Commission recently hit Meta with a record $1.3 billion fine for breaking its privacy rules. The region's new Digital Markets Act also puts some of the Meta's data-sharing and privacy practices into question.

It works a lot like Twitter

Meta bent over backward to tell us Threads is not a Twitter clone. "Threads is a new app that's focused on text and dialogue. And the way that we think about this is we're modeling it after what Instagram has done for photo and video," said Connor Hayes, a Meta product vice president.

But in many ways, Threads works exactly like Twitter. It's primarily oriented around text conversations, and your posts — called "threads" — are limited to 500 characters each. You mention other people in threads by using the @ symbol in front of their username, and can reply to someone else's posts. You can also quote or retweet — erm, "repost" — someone else's threads by clicking a button. (So should we call a thread of threads ... a knit?)

You can include photos and videos in Threads, but they don't show up as Instagram posts or Reels. You can also share Instagram photos and videos on Reels, but they show up just as regular links. There aren't native integrations between the two apps.

What's different from Twitter? There's no separate direct messaging function at launch. And you have a bit more control over the audience who can see what you post.

And unlike Twitter, Threads also has no hashtag or trending-topics function, and there's no way to edit threads once you've posted them.

You can't take your Twitter friends with you

Many Twitter users spent years curating the list of accounts to follow over on the bird app. But so far there's no easy way to port that list into Threads.

Blame a lack of interoperability. Despite the complaints of consumer advocates and some lawmakers, Big Tech companies have largely resisted calls to make their products work easily with each other. (Example No. 1: Apple's iPhone.) The companies have little incentive to make it easy for you to quit.

It's possible that someone will figure out a hack that makes this possible. Or hopefully Meta will open up technology on its side that allows you to import follower lists.

Your feed is chosen by an algorithm, not you

When you open the Threads app, you're thrown into a feed of threads that's a mix of accounts you've chosen to follow and algorithmically generated suggestions. Some of the people who are already on Threads include singer and actress Jennifer Lopez, celebrity chef Gordon Ramsay and social activist Malala Yousafzai, along with entertainment brands such as Netflix and Bravo TV.

But there's no way to make Threads show you only the posts from accounts you've chosen to follow and no way to clear the home screen of Threads from people you didn't select. In a thread, Instagram's head Adam Mosseri said a follower-only list was on the company's to-do list. Mosseri also said you can stop seeing Threads from a particular account by "muting" it, an option available behind the three-dots menu at the top right of each post.

There's also no way to make your feed ordered chronologically — instead, it's organized by what the Threads algorithm thinks you might find most interesting.

It's just as bad for your privacy as Facebook

Like Twitter, Threads accounts can be either public or private. On a public post, you can also adjust the groups who can reply to everyone, only the accounts you follow or just the ones you mention in the post.

Having said that, Threads is just as hungry for your personal data as Facebook and Instagram are. Patrick Jackson of privacy tech company Disconnect says he's found the Threads app gobbling data you might not expect, including details of your phone (model number, screen resolution and time zone) and identifiers such as a timestamp for when you installed the app.

You should also assume that when you sign up for a Threads account, Meta has access to everything Facebook and Instagram have learned about you over the years — on and off its apps — to target ads and tailor your experience.

There are no ads (so far)

At launch, Threads has no ads. But don't expect it to stay that way. Meta, which makes the vast majority of its revenue from tracking what users do online and using it to target them with ads, says it may open the door to ads in the future.

Meta says it takes safety seriously, but time will tell

Meta says safety is an advantage it has vs. Twitter. Threads will apply the same content rules that exist on Instagram. That means users on Threads aren't supposed to be able to praise terrorist or hate groups, buy firearms or make threats against people or groups.

That said, even in the first few hours of Threads, users reported examples to us of anti-LGBTQ+ rhetoric on the network — with mixed success at getting it taken down.

And keep in mind: Any user who is allowed to be on Instagram will be allowed to be on Threads. Earlier this year, Meta reinstated Donald Trump's Facebook and Instagram accounts following a two-year suspension, so he'd be welcome on Threads.

There are some protections for children. Users under 16 are automatically defaulted into a private account. And Threads will allow users to limit replies to their threads to only people that they follow or mentioned in the thread.

Threads wants to make life easier for creators

Meta says that Threads will lower the risk for creators who want to try a new text-based social media app but don't want to have to put in all the work it takes to build a new following from scratch. Instead, they can encourage their fan base on Instagram to join Threads to start. But users will still have to opt in to follow their favorite creators on Threads, and there is no guarantee that every creator's followers will be interested in the new app.

"My guess is that there's going to be a bunch of people that end up being very successful on Threads who you wouldn't really expect," Meta's Hayes said. "You might look at them now as a visual content creator, but they have a lot of great things to say they just haven't had the place to say them. And our hope is that Threads can be that place."

There's hope Threads will join the open fediverse

Meta says that it has plans to make Threads compatible with the so-called fediverse, meaning it would work along with other decentralized social networks such as Mastodon. (To do this, Meta has committed to an industry protocol called ActivityPub.) That could be revolutionary for the industry, but Silicon Valley doesn't have a great track record with actually making things work together.

Be careful before you delete it

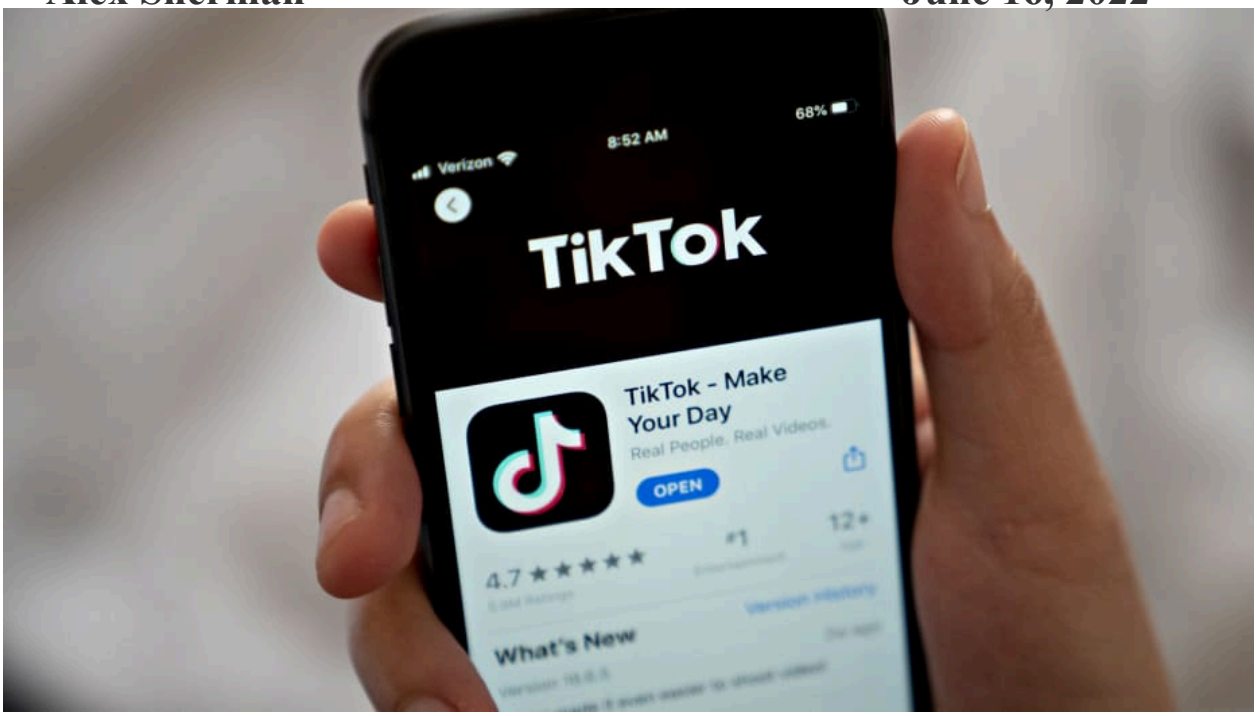
You can only delete and deactivate your Threads account by deleting your entire Instagram account.

In its privacy policy, Meta states: "You may deactivate your Threads profile at any time, but your Threads profile can only be deleted by deleting your Instagram account."

PX0318

TikTok exec: We're not a social network like Facebook, we're an entertainment platform

Alex Sherman **June 16, 2022**



TikTok is fully aware that Meta CEO Mark Zuckerberg is retooling the Facebook and Instagram apps to be more like its own popular short video service. But TikTok has no interest in mimicking Facebook.

"Facebook is a social platform," Blake Chandlee, TikTok's president of global business solutions, told CNBC in an interview on Thursday. "They've built all their algorithms based on the social graph. That is their core competency. Ours is not."

Chandlee, who spent 12 years at Facebook before joining TikTok in 2019, said his former employer will likely run into trouble if it tries to copy TikTok, and will end up offering an inferior experience to users and brands.

Facebook launched Instagram Reels in 2020 as its first real foray into the short-form video market. Last year, it brought the service over to its core Facebook app.

"We are an entertainment platform," Chandlee said. "The difference is significant. It's a massive difference."

Facebook app chief Tom Alison told The Verge this week he sees TikTok increasingly stealing share from the world's largest social network. Facebook plans to modify its primary feed to look more like TikTok by recommending more content regardless of whether it's shared by friends.

"I think the thing we probably didn't fully embrace or see is how social this format could be," Alison told The Verge.

Facebook's recent performance backs that up. Meta's stock price is down 52% this year, underperforming the Nasdaq, which has dropped 32%. In April, the company said revenue in the second quarter could drop from a year earlier for the first time ever.

Earlier in the year, Zuckerberg acknowledged the increased competitive pressure from TikTok and said, "This is why our focus on Reels is so important over the long term."

TikTok is owned by China's ByteDance, which is privately held.

Chandlee said history is not on Zuckerberg's side, and compares its current problem to the challenge that Google faced when it was trying to take on Facebook at its own game.

"You remember when Google was creating Google+," Chandlee said. At Facebook, "We had war rooms at the time. It was a big deal. Everyone was worried about it," he said.

But no matter how much money Google poured into its social-networking efforts, it couldn't compete with Facebook, which had become the default place for people to connect with friends and share photos and updates.

"It became clear Google's value was search and Facebook was really good at social," Chandlee said.

"I see the same thing now," he added. "We're really good at what we do. We bring out these cultural trends and this unique experience people have on TikTok. They're just not going to have that on Facebook unless Facebook entirely walks away from its social values, which I just don't think it will do."

Facebook didn't immediately respond to a request for comment.

Chandlee added that he has deep respect for Zuckerberg and views both Facebook and Google as strong competition. However, he noted that TikTok has an array of competitors across the world, including businesses in e-commerce and live streaming.

Chandlee said he hasn't seen a slowdown in ad spending on TikTok, despite what's being reported by companies such as Snap, which told investors that ad revenue is being hurt by inflation and the threat of recession. Snap's stock has lost almost three-quarters of its value this year.

"I've heard there's going to be a slowdown in the ad market, anywhere from 2% to 6%, but we have not seen it," Chandlee said. "We're not seeing the headwinds that some others are seeing."

WATCH: Snap has a TikTok problem, says Lead Edge Capital's Mitchell Green

PX0319

The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok

ANGELA Y. LEE* and HANNAH MIECZKOWSKI*, Department of Communication, Stanford University, USA
NICOLE B. ELLISON, School of Information, University of Michigan, USA
JEFFREY T. HANCOCK, Department of Communication, Stanford University, USA

This research examines how TikTok users conceptualize and engage with personalized algorithms on the TikTok platform. Using qualitative methods, we analyzed 24 interviews with TikTok users to explore how algorithmic personalization processes inform people's understanding of their identities as well as shape their orientation to others. Building on insights from our qualitative data and previous scholarship on algorithms and identity, we propose a novel conceptual model to understand how people think about and interact with personalized algorithmic systems. Drawing on the metaphor of crystals and their properties, the *algorithmic crystal framework* is an analytic frame that captures user understandings of how personalized algorithms (1) interact with user identity by *reflecting* user self-concepts that are both *multifaceted* and *dynamic* and (2) shape perspectives on others encountered through the algorithm, by orienting users to recognize parts of themselves *refracted* in other users and to experience ephemeral, *diffracted* connections with groups of similar others. We describe how the algorithmic crystal framework can extend theory and inform new lines of research around the implications of algorithms in self-concept development and social life.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: algorithms, self-concept, crystallized self, folk theories, TikTok

ACM Reference Format:
Angela Y. Lee, Hannah Mieczkowski, Nicole B. Ellison, and Jeffrey T. Hancock. 2022. The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok. In . ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Algorithms play influential roles in many aspects of people's everyday experiences, including their professional, entertainment, financial, and social lives. Further, people often interface with algorithms in decision-making contexts that range from the next Netflix "watch now" click, to more consequential decisions, such as whether a particular job application is passed on to a human or ignored. However, not everyone is necessarily aware of algorithmic influence in all of these contexts [27, 29]. Even when they are, people are often left to speculate with various degrees of complexity [19] about what is happening inside the algorithmic black box.

Scholars have turned their attention to studying these speculations in recent years, as well as how these speculations influence other perceptions and behaviors. This is particularly crucial, as

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CSCW '22, November 12–16, 2022, Taipei, Taiwan
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

subjective understandings "are not second best to objective measures; in fact, it is perceptions and subjective realities that people judge and act upon" [67]. Perceptions of algorithms can be shaped by individuals' lay theories about their functioning and knowledge about the technical processes that underpin personalization (e.g., algorithmic literacy) [25, 43]. Further, because algorithms change as a result of the data input they receive in a kind of feedback loop [50, 55], Bucher [9] even argues that these perceptions can then influence the nature of the algorithms themselves. Indeed, Karizat et al. [38] refer to this phenomenon as a process of co-production, in which people may influence algorithmic output, but algorithmic output may also influence people's perceptions.

A great deal of scholarly attention has been paid to TikTok in particular as the platform has grown in popularity (e.g., [2, 5, 40, 65]). The For You Page (FYP) is of particular relevance, as it is the first feature of the platform that users interface with, and involves algorithmically-curated videos purportedly based on users' interests. In the current study, we aim to understand how TikTok users think about the FYP algorithm as it relates to their identity. We focus on the idea of self-concepts [44] to describe the many identities, roles, beliefs, and values that compose an individual.

In addition to potentially emphasizing relationships between perceptions of the algorithm and perceptions of the self, TikTok also affords strangers as the main audience for one's content [4]. Instead of direct interactions with networks of known ties, Zulli and Zulli [77], as well as Abidin [2], suggest that social media platforms like TikTok afford a different kind of social experience. As such, people's self-concepts may be particularly salient in terms of how they perceive others on the platform. As such, we ask the following:

RQ: How do participants perceive the algorithm in relation to their self-concept?

We conducted semi-structured interviews with 24 self-reported active TikTok users. The results of this study suggest that our participants recognized parts of their identity in the FYP algorithm's content recommendations. Subsequently, they felt a desire to bring algorithmic representations of themselves into alignment with their own self-concepts. Participants felt that their behaviors could shape the algorithm's ability to accurately reflect their multifaceted and dynamic interests and preferences. Further, the algorithm facilitated feelings of "mere belonging" towards others on the platform, with multiple participants reporting positive feelings toward others even though little to no interpersonal interaction took place. We synthesize our results with prior literature on algorithmic interactions and self-concepts, and introduce an analytic framework - the *algorithmic crystal framework* - that coalesces our findings and can help guide future work.

1.1 Related Work

One way to approach investigating perceptions of algorithms and the sociotechnical systems in which they operate is by exploring algorithmic folk theories, or "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems, which guide reactions to and behavior towards said systems" [21]. Although these folk theories are developed with varying levels of awareness of algorithmic capabilities [19], they can shape people's behaviors on platforms, which in turn can be treated as input data for the algorithmic decisions themselves. A dominant algorithmic folk theory is that of Personal Engagement [25], in which people anticipate that the more they interact with a specific type of content, the more they would see similar content on their algorithmically-curated feed. In a general sense, many algorithmic systems that determine the content that users are exposed to actually work in this way as well [62]. Overall, algorithmic folk theories can be considered a collection of perceptions about algorithms that influence peoples' behavior when interacting with these systems.

Other scholars have investigated perceptions of algorithms from different lenses [52, 53]. For instance, Bishop [7] examined "algorithmic gossip" or how people talk about their algorithmic interactions with others. This is especially prominent among content creators; through this discourse, content creators may change their behaviors in order to achieve goals like gaining visibility on a certain platform. Relatedly, Siles et al. [56] also approaches perceptions of algorithms from a more pluralistic sense. These authors argue that it is through people's collective algorithmic imaginaries, or "ways of thinking about what algorithms are, what they should be, how they function and what these imaginations in turn make possible," [9] that large scale changes in human-algorithm interactions can be enacted. Although these concepts highlight various mechanisms through which scholars can better understand interactions between humans and algorithms, they converge by focusing on perceptions of these technologies, and how these perceptions may influence behavior.

In particular, the role of identity in algorithmically-curated interactions has been a prominent area of investigation, especially on the TikTok platform (e.g., [6, 38, 41, 59]). For instance, some TikTok users perceive that the "algorithm recogniz[es], classifi[es], sort[s], and suppress[es] social identities" on the platform, replicating extant gender-based or racialized power dynamics present outside of TikTok [38]. Indeed, popular media has also reported on user experiences related to thinking about the self, speculating that the videos shown to users say something about the type of person they are [68], or even allow users to better understand their own identity [8, 45].

Researchers have often examined self-concepts - or the the many identities, roles, beliefs, and values that compose an individual [44] - on social media as they are demonstrated through self-presentation (e.g., [12, 64, 72]). Recently, Simpson and Seeman [59] have investigated how algorithmically-curated content can both affirm and violate notions of one's own identity, especially in the case of users with marginalized identities (e.g., LGBTQ+ users). French [28] also found similar effects, even indicating that people's perceptions of themselves may change based on information provided by an algorithm. In the case of TikTok, Bhandari and Bimo [6] argue that because users overwhelmingly engage with a "personalized algorithm which repeatedly confronts users with various aspects of their own personas" called the For You Page, the processes of self-perception may interact with platform affordances in unique ways and rely on different mechanisms than in previous research on the topic.

Since the affordances on a platform can allow, or even encourage or discourage, various behaviors [18], people's experiences on TikTok may be shaped by what features are or are not available. Even though other platforms such as Twitter or Reddit allow interactions with strangers, for instance, the visual salience of the content creator is not highlighted in the same way it is on video-based platforms like TikTok [42, 77]. Further, the aforementioned algorithmically-curated FYP is arguably the most salient feature on the platform. As such, "network form[ation] through processes of imitation and replication" [77] in terms of visual content is common, meaning that users do not necessarily need to have direct social interactions with others to form connections. The networks of content connected by similar algorithmically-curated content on TikTok have been discussed as "rabbit holes," "silos" and "subcultures" [1], as well as "sides" [24]. TikTok does allow the formation of interpersonal networks with the ability to follow one's known ties or direct-message others. However, predominance of the FYP, combined with the visual salience of the content creators, is in notable contrast to many popular social media platforms which rely heavily on content from known ties in a network, rely less on visual content, or both.

The lack of direct, interpersonal connections between platform users may mean that how people think about others, specifically in relation to themselves, has the potential to be particularly salient. Further, the ways in which algorithmically-curated content interacts with perceptions of the self may be especially notable as well, as recent researchers have also noticed. In this paper we examine

how TikTok users perceive the FYP algorithm, particularly in the context of personalized content, as well as how perceptions of the algorithm influence beliefs about the self.

2 METHOD

2.1 Participants

Our sample consisted of 24 students recruited from three universities in California, two community colleges (n = 13) and one private university (n = 11), in order to increase the potential for diversity in our sample. They were recruited through an online participant recruitment system called Sona to take part in a "TikTok Study." Sona is an online scheduling service used by the universities to manage their research participant pools. By signing up to take part in our study on Sona, students attending these universities can complete research studies in exchange for research credits, which are a required component of their social science classes. We stopped recruiting participants after achieving saturation, or when we started hearing similar ideas from participants [15]. Participants were required to be 18 years or older and active TikTok users in order to be eligible to participate. We allowed participants to interpret the term "active" however they wished. The average age of participants was 23 years old (range: 18 - 43). The majority of participants identified as female (n = 18; male, n = 6). Our sample included 11 participants identifying as Hispanic/Latino, 8 as White, 4 as Native American/Indigenous, 4 as Asian/South Asian, and 1 as Black/African American, with some identifying with multiple identities. The majority of participants were heterosexual (73.9%), with 1 identifying as queer, 1 as bisexual/pansexual, 1 as bisexual, 1 as undecided, and 1 who declined to respond. In order to protect the privacy of participants in our relatively small participant pool, we report the demographic information in aggregate above, as opposed to in an individualized manner in Table 1. All participants received research credit, part of a course requirement, for their participation.

Table 1. Self-reported participant demographic information

Participant	Age	Gender	Participant	Age	Gender
P1	22	Male	P13	19	Female
P2	22	Male	P14	20	Female
P3	21	Female	P15	18	Female
P4	21	Female	P16	18	Male
P5	20	Female	P17	24	Female
P6	18	Male	P18	25	Female
P7	19	Female	P19	30	Female
P8	24	Female	P20	22	Female
P9	20	Female	P21	43	Male
P10	20	Female	P22	31	Male
P11	21	Female	P23	24	Female
P12	31	Female	P24	27	Female

2.2 Procedure

Data collection consisted of interviews that were conducted remotely using the video-conferencing platform Zoom due to the COVID-19 pandemic. The interview protocol was developed by the research team in order to provide opportunities to explore users' perceptions, engagement, and reasoning about the TikTok platform. All interviews were conducted in a semi-structured format

to allow researchers to explore and probe participants' perspectives. The final protocol consisted of 32 open-ended questions that asked about participants' perceptions and attitudes about the TikTok app (e.g., How would you describe TikTok? How do you tend to use it?), their understanding of the different types of content (How would you describe the content you see on TikTok?), and how they understood the functioning and characteristics of the algorithm (How would you describe the TikTok algorithm if it were a person?). We also asked participants to report demographic information about their age, gender, race and ethnicity, and sexual orientation at the end of the interview, with all participants having the option to not answer any questions. This study was approved by the IRB at the first authors' university.

2.3 Research Positionality Statement

Some of the identities represented in our research team included woman of color, immigrant, and first-generation college student. Our team includes experts in social media, algorithms, and folk theorization.

2.4 Data Analysis

Interviews were conducted remotely by the first authors from February through May 2021. The interviews were recorded using the record-to-computer feature on Zoom with participants' consent, and the audio was transcribed using an automated online transcription service (Otter.ai). The first authors removed identifying information from the transcripts, assigned all participants an ID number (e.g., P1, P2), and flagged instances of technical errors in the recordings (e.g., from poor Internet connection). All transcripts were checked for accuracy twice, first by one of four research assistants and then by the first authors. The checked transcripts were imported into the qualitative coding software ATLAS.ti for analysis and coding. Quotes included in the manuscript were lightly edited for clarity.

The first authors generated an initial list of in-vivo codes, using phrases from participants themselves (e.g., the "relatable" quality of other users) which were iteratively reviewed and discussed by the research team. Example codes included feeling "seen" by the algorithm, feelings of control over the algorithm, and the diversity or similarity of content identified by the algorithm.

New axial codes were added to the codebook throughout the coding process to capture new insights and to identify higher-order themes (e.g., strategic engagement with the algorithm) [73]. Example codes included lay or folk theories about the algorithm's functioning, perceptions of the algorithm (e.g., as a person), and signals of algorithmic personalization. Throughout this process, we wrote research analysis memos [46] where we identified key themes that emerged from the qualitative coding and review process, exemplar quotes, and questions for discussion with the research team. Materials including the full interview protocol and list of codes are available upon request from the corresponding author.

In addition, we created a data matrix to identify higher-level patterns at the participant level [46]. For example, we classified participants by their degree of reported perceived personalization of the TikTok For You algorithm, their sense of belonging or engagement with TikTok groups or "sides," the amount or intensity of their TikTok use, and whether they created content for TikTok or predominantly consumed it. Using this matrix, we were able to organize our observations of individuals' perspectives on the algorithm around individual differences, such as users' preferred method of engagement or the amount of their use.

3 FINDINGS

3.1 The Algorithm Knows Me Well: Reflecting Multifaceted and Dynamic Self-Concepts

We examined how individuals theorized about the relationship between the algorithm and their sense of self-concept. As other research has observed (e.g.,[25, 57]), some participants evaluated the algorithm more positively when it offered them relevant or personalized content. The majority of participants, particularly those who were heavy or frequent TikTok users, believed that the algorithm learned who they were and what they liked by analyzing traces of their behavior. Prior work has discussed this kind of theorizing as an algorithmic folk theory of personal engagement [25]. While some participants expressed uncertainty and curiosity about how the algorithm worked, many perceived their experiences as the algorithm attempting to understand and reflect them.

Specifically, some participants emphasized the importance of the algorithm's ability to recommend content in line with a sense of self-concept that they viewed as fundamentally *multifaceted* and *dynamic* - two characteristics we discuss in turn below. Based on our analyses, we found that participants noted that they had varying identities and interests, some of which remained relatively stable and constant over time (nationality, native language, a life-long passion for theater, etc.), and others that were more likely to evolve or change (photography, nature, baking, or a particular TV show, etc.). How good an algorithm was perceived to be, then, was a function of its sensitivity: the algorithm's continuous "tailor[ing]" or "mold[ing]" to the set of facets that comprised users' self-concept (P4, P10, P15), resulting in a feeling one participant described as the algorithm "know[ing them] so well" (P10).

Many tended to value moments when the algorithm seemed to reflect many different facets of themselves. Rather than providing recommendations for only one or two key interests, the TikTok algorithm was described as "tailored to everything you'd want to see and be entertained by" (P14) and "showing me everything, like everything from the corners of my mind" (P8). For example, one participant (P13) enthusiastically attributed her enjoyment of TikTok to the diversity of content recommendations that spoke to her many specific interests, commenting that "it's nice that they all blend into [a] customized page for me." Having described herself as being passionate about education, pets, theater, horse videos and Vine-like humor, among other topics, she said:

> *"It doesn't take long for [TikTok] to calculate what videos I watch, what videos I don't watch, what my engagement is. So it just becomes more and more specific the more I use it. I don't see how it can get any more specific, like it's pretty insane how accurate it is right now."*

As demonstrated by the above quote, the specificity of the algorithm was determined by its ability to aggregate and recommend content that reflected the many and multiple facets of participants' self-concept. The "insane" accuracy of the algorithm, then, was about creating a unified blend of customized content recommendations that reflected users' diverse interests and identities into a single stream. It was not enough for the algorithm to provide highly specific recommendations for one particular interest, which could be interpreted as the algorithm producing a reductive representation of the self. For example, one participant (P14) preferred TikTok's algorithm over the Instagram Explore page because they thought it was better at capturing the multifaceted nature of their personal interests:

> *"With the Instagram Explore page, I feel like there's so much less variety, like if you click on a specific kind of post more than [once], it'll just give you a lot of that same post. Whereas with TikTok, I feel like the selection is so much larger... Even while tracking what you're more likely to enjoy or view… it's still a much wider selection of different kinds of content, or different kinds of styles of posting or different trends."*

In addition to being multifaceted, we observed that most participants tended to view themselves and their interests as changing over time - or as being dynamic. Thus, they wanted the algorithm to be responsive and rapidly update its representation of the user's self-concept. For example, the TikTok algorithm was seen as better than others that were perceived to be more stagnant, which caused some people to feel constrained by older identity elements and outdated preferences and thus typecast as one-dimensional. For example, one participant (P13) described TikTok as being able to "catch on to what I like, almost immediately" whereas the Instagram algorithm "took a long time to catch up to what I like":

> *"[My Instagram Explore Page] has not updated in the slightest, no matter how many times I will 'like' different things or whatever. I had a huge Disney phase, it's all Disney stuff on my Explore page, and I'm so sick of it. So just that fact alone is making me use it even less and less."*

In this quote, we can see that the participant felt "sick of" the Instagram Explore page algorithm because it was recommending content for an outdated version of herself that she no longer identified with. Being shown a reflection of her former self as an avid 16-year old Disney fan by the algorithm felt uncomfortable and frustrating. As demonstrated in the above example, the discontinuity between one's current self-concept and an algorithmic representation of a previous self can create a sense of temporal dissonance in identity, especially when differences between past and present selves are salient - which might be particularly relevant for emerging adults [3] transitioning from Disney-fan childhood to adulthood.

However, this does not mean that participants wanted the algorithm to continuously start from scratch when attempting to reflect their current interests. A common theme we observed across participants was a tension between wanting the algorithm to cater to past interests, while also being receptive to new interests and creating opportunities for personal growth and new discoveries. Although some participants expressed that they did not want the algorithm to concentrate on specific interests, others pointed out that they valued when the algorithm was able to remember former interests and rotate through a combination of former and newer interests in a way that they described as "time-shifting" (P2). As one participant explained, she felt that her experience with TikTok involved "never staying on one [interest area] for long," but rather that her feed "goes through a cycle" (P13). For example, she appreciated when the algorithm appeared to remember that she enjoyed watching videos from Frog TikTok; even if frog videos had not "come up for a while" on her feed, she felt confident that it would come back up again eventually. Another participant (P2) voiced a similar sentiment, saying:

> *"And I think that kind of goes back to the algorithm too… There's a lot of people that like to talk about a specific thing, and that obviously shifts over time. But there are certain communities probably that persist over time, whether it's about elections... there'll always be that community of posts that you can probably find if you like something or look at a profile that has something like that."*

Together, these findings suggest that people understood personalization processes in the FYP as the algorithm trying to develop an accurate representation of their self-concept, preferably in a way that captured their multifaceted interests and dynamic ability to change over time. At a high level, we observed that some users believed the algorithm could know their "their true self" because it was capable of identifying content in line with the contours of their identity.

3.2 Aligning the Algorithmic Self with the Actual Self

In addition to evaluating how well the algorithm's representation of them fit with their self-concept, almost all of our participants had a working understanding that the algorithm's representation of

them was shaped by their everyday use of the app, such as the videos they watched and the content they engaged with. Whereas previous work on algorithmic selves has highlighted that there are various degrees of user awareness of how their data are represented through datafied systems [11], our findings indicate the potential for users to not only be aware that this personalization process is occurring, but also to want to bring their algorithmic self into alignment with their actual self.

One way some participants thought they could shape the algorithm was casual and unintentional: simply by being themselves and following their natural interests when engaging with content. Many reasoned that the algorithm could learn who they really were over time because it was making inferences from their everyday watching behavior, such as the number of times they watched the same video, whether or not they investigated a particular content creator, or if they skipped over a certain type of content ("I think I definitely enable it when I watch a video maybe twice... or if I look at the comments, or if I like it" - P5). Unlike other social media platforms with stronger norms around engaging with content to maintain social ties (e.g., feeling obligated to "like" a friend's Instagram pictures [32]), the FYP algorithm could build an understanding of the user's interests through content they organically gravitated to. While some participants saw this as producing quick, easy, and accurate representations of themselves, others felt like they had little control over what the algorithm learned in this fashion because it was like it was "read[ing] your mind" (P15). As one participant expressed, "It's like he knows everything… the thing that we cannot control is definitely the AI" (P19). Thus, in this process, the algorithm's representation of the self could be closely aligned with one's actual self with little effort on the part of the user, but also difficult to evade or fabricate.

In contrast to these organic expressions of interest, another way that participants attempted to align the algorithmic self with their self-concept was more effortful and strategic. Some participants described feeling like they could shape what the algorithm learned by actively curating their engagement through deliberate interactions with the app. In this process, algorithms were valued when seen as responsive to bids from users to adjust its representations of them and recommendations for them. For example, several people described how these bids could be made explicitly by deliberately engaging with features to send a request to the algorithm (e.g., indicating they're 'not interested' by clicking on this option or saving the video):

> *"If I really liked the video, and I want similar ones, it's obviously liking the video or commenting, or adding it to my [collection of Favorited videos]. ... I know for videos [that] I don't want to see... you can press the video and it'll pop up like, a 'not-interested' option... Sometimes I'm like, 'No, I don't want to watch this.'"* - P9

Users could also make bids implicitly by modulating their behavior in hopes that the algorithm would pick up on this change and thus learn to better model some facet of themselves. For example, in the context of describing her feed, one white participant (P18) described how she noticed that the algorithm was showing her videos of predominantly white creators and that she tried to deliberately change this by strategically engaging with Black creators:

> *"Yeah I would say that at first I was probably seeing mostly white people... But again with BLM and everything - I started liking content by more Black creators and a more diverse set of creators."*

We observed that people tended to make different attributions when they made a bid to the algorithm to change how it represented them. Some people felt they needed to do so because the algorithm was not sensitive enough to who they were, or that it was biased in its own way to showing particular content - which reflected poorly on their evaluation of the algorithm. A few participants, however, expressed a tension between their actual behavior and the kind of person they wanted to be. In this frame, they viewed the algorithm itself as highly capable of learning and modeling who they

were, but felt uncomfortable about some of the identity facets that it brought to light. For example, this tended to involve content recommendations that indulged in interests that they viewed as unhealthy, such as snack foods. Seeing unsavory facets of oneself detected by the algorithm and reflected in its recommendations could spark a potentially discomforting process of self-reflection and awareness.

Rather than viewing the algorithm as a purely mechanical function that recommended relevant content, we found that participants tended to anthropomorphize the algorithm to varying degrees. Some made sense of the personalization processes underlying the algorithm as being a "people-pleaser" (P22) that was capable of changing itself to be as similar to the user as possible. As one participant (P6) explained:

> *"It's whoever you are. It's like... the Pokemon that just turns into you. Or tries to. I mean, my TikTok experience is going to be completely different from someone else's because [it] got so personalized to me so quickly... So the algorithm is just you. Or it's just trying to be you."*

For participants like P6, the algorithm was viewed as a fundamentally malleable entity whose core mission is to transform itself into a copy of the user by learning to represent their interests and identities. However, people varied in how they thought about the goals of the algorithm in trying to learn and satisfy their tastes. For example, some participants viewed this in a purist, positive light, characterizing the algorithm as someone who enjoyed meeting their needs: "It would be someone who's very obedient and very, like a people-pleaser. They like to accomodate for you" (P15). In contrast, others were more critical, viewing the algorithm's motives for wanting to please them with more cynicism or apprehension. For example, one participant framed the algorithm as ensnaring, describing it as someone who "like, just wants to give you whatever you want" as a ploy to "keep you there [on the app] and not let you go" (P11). On the other hand, participants like P7 viewed the algorithm's people-pleasing qualities as less malicious, but still hollow or insincere:

> *"It's kind of like, if someone were to change their personality with everybody that they met in order to be - not necessarily the best person with the other person, but what they expect the other person would want, who they would want to hang out with."*

Despite these differing perceptions, however, people's understanding of the algorithm appeared to be unified by a shared orientation to viewing the algorithm as attempting to represent themselves, with more or less fidelity. Overall, our findings suggest that people's understanding of personalization processes shaped how they thought about the algorithm's representation of themselves, why it was learning to form such a representation, and how to bring their algorithmic self and actual self-concept into alignment.

3.3 I'm Not the Only One: Algorithmic Exposure to Others on the Platform

Participants were able to experience a variety of sides, or subcultures [1, 24, 77] on their own algorithmically personalized feed, which contributed to an acknowledgement of themselves as complex, multifaceted, and dynamic. In comparison, they often had a more simplified, one-dimensional perception of others who happened to be on the same side as them. Rather than interacting with others on the platform through a conventional network of known ties and with direct interpersonal contact - like dyadic or group messaging on Facebook - our analyses suggested that people tended to view others' content through the lens of their own self-concepts as curated through algorithmic personalization.

Perhaps the most consistent perception mentioned by participants was the relatability of content creators on TikTok, regardless of what category of video they made. One participant (P3) mentioned that they were seeing people "very similar" to them, while another (P12) found that she enjoyed

"watching stuff that [she could] relate to about parenting" because of the parts of participants' self-concepts that overlapped with content creators. Other participants remembered thinking "oh my god, I'm not the only one who feels this way!" (P23) after seeing content about a TV show they enjoyed or that they were "not alone" (P24).

In contrast with other platforms that have algorithmic recommendation systems that primarily rely on content from members of one's network of known ties, it was not considered particularly important for participants on TikTok to have many overlapping interests with a singular other user on the platform. For example, one participant (P18) who enjoys musical theater and singing noted the following about Kristen Chenoweth, an actress and singer:

> *"She posted another video where she was harmonizing with the fire alarm… I've had similar things where... I'm like, 'Oh... the vibration from my toothbrush makes this sound… and then lets me harmonize to it.' And so it's totally relatable."*

Where does this feeling of relatability stem from, even among participants and content creators who have minimal interests or identities in common? In the case of our participants, relatability appears to be driven by the familiarity of another's self-concept when compared to one's own, even if these self-concept dimensions are only partially overlapping. Further, relatability is largely experienced in a content-based, rather than person-based manner, similar to the phenomena noted by Zulli and Zulli [77], as well as Abidin [1] and Klug et al. [42]. One participant (P6) even mentioned that people were just "vessel[s]" for relevant content that could have "anybody's face" on it, exemplifying the relative lack of importance of personal relationships as opposed to personalized content.

To further explain this phenomenon, we draw on Walton et al.'s [74] proposition of mere belonging, or the "minimal, even chance, trivial, or potential, social connection with unfamiliar others." In past psychological research, feelings of mere belonging have led to increased interpersonal liking or better outcomes in social interactions, even in transitory or short-term settings. However, many of our participants had minimal direct social interaction with content creators or other users of the platform, although exposure to videos remained transitory. Some participants were even able to pinpoint the source of mere belonging, with one participant (P7) stating that "you don't talk to [people on the sides]... you don't know them, but you feel that sense of connectedness because you both… like one thing." A noteworthy instance of this was described by another participant (P23):

> *"I am Mexican, but it was very hard for me to learn Spanish. And so TikTok made this whole thing - they call them the 'no sabo' kids. [It] means 'they don't know' Spanish. So I was on 'no sabo' kids for a long time and they just don't know Spanish [either]. And it was cool to see how many people who are so into their culture and love everything about being Mexican - but just don't know how to speak Spanish - come together. It was really nice actually."*

In the case of this participant, no sabo kids are people who share a specific life experience with her, which in turn brings about positive feelings towards them, but who she otherwise does not directly interact with. Other participants noted similar feelings of ephemeral belonging to the sides of TikTok that they were also on. One participant (P2) mentioned that she "never post[s] videos" and she "very rarely comment[s]" but does feel that other TikTok users "share similar sentiments as [her]." For these participants, algorithmically personalized content allows them to see others who share similar aspects of their self-concept and reap some social benefits, such as feelings of connection, without direct interpersonal interaction.

Seeing others' content through the perspective of parts of one's self-concept, as well as the effect of mere belonging, can be further accentuated by the presumed number of people on a side. Participants described the benefits of sides with fewer other viewers, such as feeling more unique. One participant (P16) noted that "most of the time…I see [the videos] before [they] blow up."

Similarly, another participant (P6) noted that "once you get deep enough into the TikTok [sides], you get a lot of videos that don't have that many views so you feel like you're… really important." Further, P9, who creates content, mentioned that they prefer "a smaller audience... so I can actually see... how [people are] relating to the video."

Although there is no way for participants to truly know how many people are on the same side they are on, they can calculate estimates using the metrics that TikTok makes visible, such as the number of views, likes or comments. Overall, participants intuit that greater degrees of algorithmic personalization, experienced through hyper-specific content made by others, means that aspect of their self-concept is potentially more unique.

4 DISCUSSION

This research explores how algorithmic personalization processes inform user understandings of identity and shape their orientation to others. We focus on the platform TikTok, widely recognized colloquially for the responsiveness of its content recommendation engine [61, 63], and explore how TikTok users perceived the FYP algorithm in relation to their self-concept. Our participants' experiences with the FYP algorithm reflected a recognition of parts of their identity in the content recommendations and a subsequent desire to bring algorithmic representations of themselves into alignment with their own self-concept. Our participants generally understood this alignment to be a cyclical interaction of co-production between the algorithm itself and their own actions [38], whereby the algorithm was appraised in terms of its ability to reflect the multifaceted and dynamic interests of the user. At the same time, our participants felt that their behaviors could shape the algorithm's ability to accurately reflect their interests and preferences. Shaping of the algorithm happened in two distinct ways: through participants' organic interactions with platform content and through their strategic attempts to influence the algorithmically-curated content.

Despite tending to just watch content on the platform rather than creating content or engaging directly with other users, multiple participants reported positive feelings toward other people who had similar interests or identities - even though little to no interaction took place. Other people shown to them on the platform were seen as highly "relatable," because they were understood to be similar to the user in some way deemed relevant by the algorithm. As such, the algorithm facilitated feelings of "mere belonging" towards others on the same side as participants.

We synthesize and extend these findings below by introducing an analytic framework that coalesces our findings and can help guide future work, drawing from literature on algorithmic interactions and self-concept. This framework contributes a new approach to understanding how people understand and experience personalized algorithms.

4.1 The Algorithmic Crystal Framework

Building on insights from our qualitative data, as well as previous scholarship on the "crystallized self" [70, 71], how people perceive personalized algorithms [19, 25, 28], and relationships between the TikTok algorithm and identity (e.g., [6, 38, 41, 59]), we propose a novel conceptual model to understand how people think about and interact with personalized algorithmic systems. Drawing on the metaphor of crystals and their properties, the *algorithmic crystal framework* is an analytic frame that captures user understandings of how personalized algorithms (1) interact with user identity by *reflecting* user self-concepts that are both *multifaceted* and *dynamic* and (2) shape perspectives on others encountered through the algorithm, by orienting users to recognize parts of themselves *refracted* in other users and to experience ephemeral, *diffracted* connections with groups of similar others. We discuss each of these properties in detail below.

4.1.1 Reflective. Crystals have several properties that we draw on for our framework. Crystals are solid structures consisting of flat facets which can be shaped over time into multidimensional prisms that capture and transform light. One transformation is reflection, or returning light back to its source. In the algorithmic crystal framework, algorithms are understood to be *reflective* of various dimensions of the self, including one's interests and identities. This echoes French's [28] notion of the algorithmic mirror, in which algorithmic systems "mine one's online activity, infer one's traits, and reflect those traits back to a person in the form of personalization." Similarly, Hess [34] also argues that interactions with digital technology are often "more like looking into a mirror than looking out a window" in that we see more parts of ourselves than we see of others. Further, Seaver [54] describes how successful recommender systems have to work in conjunction with users' psychological orientations in order to be truly captivating. In our data, reflection took the form of the FYP algorithm curating content for participants that they felt represented their interests or identity, supporting similar findings in other research on TikTok [4, 37, 57]. That is, people felt that parts of their self-concept were accurately *reflected* back to them in the content provided by the algorithm.

4.1.2 Multifaceted. The algorithmic crystal framework reminds us that algorithmic reflections of the self can be *multifaceted* (see Figure 1) as well, as opposed to being flat, one-dimensional, or oversimplified. Similar to work on the crystallized self, which draws on the imagery of the crystal as prismatic to characterize the self as fundamentally multidimensional [70, 71], we found that our participants valued when the algorithm was able to reflect their self-concepts as a constellation of identity facets. These ranged from core traits like ethnicity and sexual orientation and deeply-held passions, to fleeting interests and new discoveries.

Previous work, particularly on TikTok, has found that people understand personalization processes as placing people into different categories and showing content that is thought to fall within these categories [57]. Indeed, one complaint that participants often have of personalized recommendations is that "maybe someone isn't defined by a [single] category" [26]. Karizat et al. [38] also notes the potential expansiveness of one's self-concept in a related way, discussing the reflections of both person and social identities. A crystal can be held up to the light and rotated to reveal one facet at a time; likewise, our participants were shown one piece of content at a time on their FYP and were able to recognize the various facets, or "sides," that pieces of content belonged to. The metaphor of the algorithmic crystal, then, highlights how algorithmic representations can capture and reflect users' many identity components but also, when rotated, allow users to discover new ones. Just as elements of identity can range from the general to the niche, facets can come in a variety of sizes.

4.1.3 Dynamic. The algorithmic crystal framework also highlights how the self is *dynamic* and changes over time. Though the hard structure of crystals may evoke images of rigidity, prior work on the crystallized self highlights that they can take on an infinite variety of shapes, that they can "grow, change, alter, but not [be] amorphous" (Richardson in [71]); with the algorithmic crystal, this fluidity is reflected back to the user via content recommendations that similarly adapt and shift over time (see Figure 1). Just as users may grow in new directions and explore new interests, the algorithmic crystal is assumed to be continuously changing: creating, removing, and refining new facets to represent the user over time. The dynamic nature of the algorithmic crystal also emphasizes the role of the user in shaping the form of their algorithmic representations, a process that Karizat et al. [38] discuss as a form of co-production: wherein individuals shape the algorithm's recommendations and the algorithm in turn shapes how they understand themselves.

Other research on TikTok has highlighted the algorithm's dynamic nature as well; Simpson and Semaan [59] claim that "engaging with content by liking it… was mentioned by all of [their]

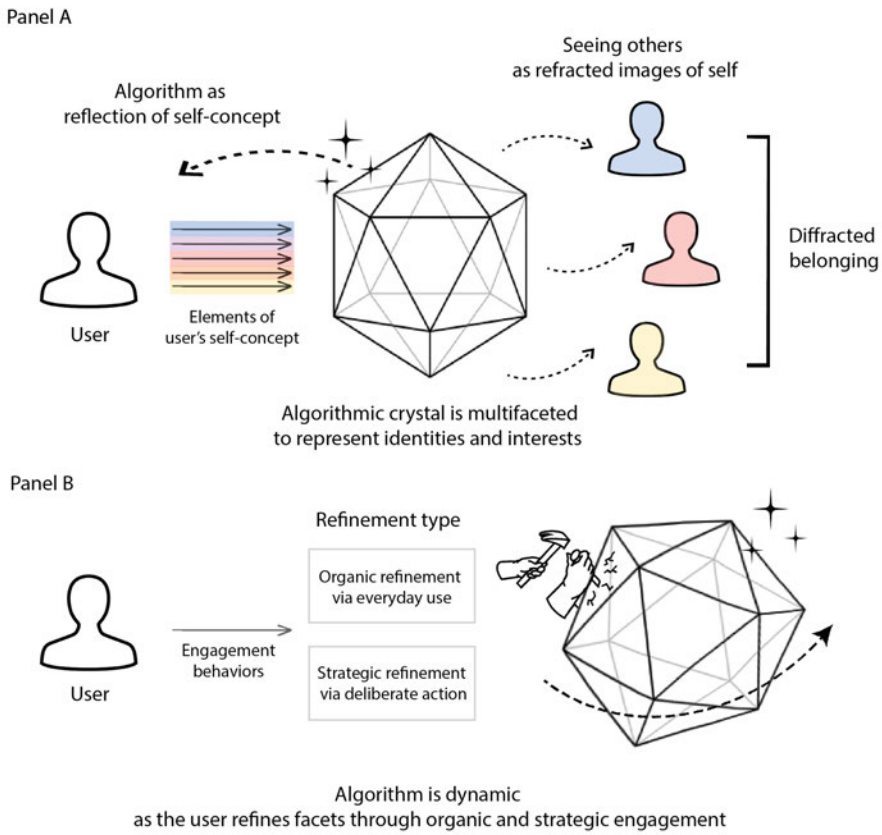


Fig. 1. A visual representation of the algorithmic crystal framework in two panels. Panel A emphasizes the reflective and multifaceted nature of the crystal, as well as the ability to see refracted images of the self and feel diffracted belonging. Panel B emphasizes the dynamic nature of the crystal and types of refinement.

participants as a way to influence the FYP algorithm over time," with other scholars noticing similar patterns of behavior [38, 57]. Just as crystals can be processed to better reveal or occlude underlying patterns - cleaved, polished, even bruised - the algorithmic crystal can be transformed by the user to better reflect their own sense of self, whereby their algorithmic self and self-concept are aligned. We understand this alignment to be desired by users, for both pragmatic (i.e., better content recommendations) and psychological reasons. For instance, people often experience positive feelings about the self derived from seeing two versions of oneself in harmony, rather than dissonance [14, 36]. Further, recent empirical work has indicated that people may feel like certain technologies are extensions of the self [17, 48, 51]. Indeed, Ross and Bayer [51] note that technologies may be perceived as "a part of the self or... a reflection of the self."

4.1.4 Refinement Strategies. People can dynamically refine the facets of the algorithmic crystal presented to them by engaging with the platform and providing more information about their preferences and tastes; this can occur organically or strategically. *Organic* refinement happens naturally, when users interact with content without the conscious goal of training the algorithm. Watching a particularly engaging clip two or three times, or saving it to show a friend later, would

fall into this category. Siles González and Meléndez Moran [57] conceptualized similar interactions between users and the TikTok FYP algorithm as a "depuration" process, in which "impurities" are removed from their feeds.

In contrast, *strategic* refinement, or polishing, describes deliberate and effortful attempts to train the algorithm with the goal of eliciting content that is more aligned with one's current or desired self-concept (see Figure 1). People may vary in the extent to which they feel that the algorithm is reflective of who they truly are. However, they may also employ a variety of strategies to try to bring the algorithm's recommendations into line with who they are, or want to be. Participant P18, who consciously sought out non-White content creators, engaged in a form of strategic algorithmic refinement. This refinement can be understood as a form of labor, where certain individuals exert time and energy to make the algorithm their own [38, 58]. There may also be elements of someone's self-concept that they do not want an algorithm to recognize, such as an identity they are still exploring or an interest they are hoping to move away from. As a result, they may employ refinement strategies to shape the algorithm such that it reflects the person they would like to be.

As work on algorithmic domestication suggests [58], however, individuals may vary in the perceived amount of control they have over the algorithm and its representation of the self. For instance, not all users conceptualize algorithms as entities to be strategically polished or have, as described by Devito [19], high folk theorization complexity levels that support them in shaping the algorithm. However, refinement offers an important pathway to conceptualizing individual agency in the context of algorithmic systems [39]. Users can exert influence over what they see and how the algorithm sees them by purposefully avoiding clicking on content they do not want to see more of [22] or by purposely engaging with creators of marginalized backgrounds [38], like in the case of P18 in our dataset. It is important to note that organic and strategic polishing may look the same, because they involve the same behaviors, but motivations for the behaviors are quite different. Distinguishing between these two forms of refinement allows us to acknowledge the role of individual intentions in refining algorithmic representations of the self.

4.1.5 Refractive. Crystals can also bend the direction of light through refraction. Similar to reflection, the process of *refraction* highlights perceived or presumed similarities between the self and other users, such as content creators. Viewing the algorithm's recommendations as a reflection of who they are and what they like, people can come to see others that they encounter through the algorithm as *refracted images* of themselves: an egocentric orientation where people recognize at least one salient component of their own self-concept in others. In the case of our participants and the FYP algorithm, this might mean they recognize a sense of humor, a love of mystery books, or a shared ethnic or professional identity. As explored in other work on the experiences of LGBTQ+ users, finding facets of one's identity represented in algorithmically-shaped spaces like the LGBTQ+ sides of TikTok can help people feel seen and affirmed in that aspect of their identity [59].

In Tracy and Trethewey [71], Richardson also notes that "crystals are prisms that *reflect* externalities and *refract* within themselves" [emphasis added]. In the same way that external images viewed through a crystal are often warped by the contours of its internal structure, seeing another person through the algorithmic crystal can produce an altered view that emphasizes shared facets of interests and identity. Indeed, people can understand the algorithmized self as engaging reflexively with previous representations of themselves [6]. As a result, other users encountered through personalized algorithmic content may feel particularly relatable even in the absence of traditional markers of interpersonal communication and relationship development, such as message exchanges, because they feel like a refraction of the self. The salience of visual content and lack of direct, interpersonal interactions on TikTok, as well as folk theorization about how algorithms organize content around interest facets shared by many people, likely contribute to this process of refraction.

4.1.6 Diffractive. Lastly, crystals can split beams of white light into a spectrum of colors through diffraction. The algorithmic crystal can also *diffract* individuals' self-concepts into its constituent parts in much the same way as a prism breaks up a singular beam of light into a spectrum of colors. Unlike algorithms that attempt to find another individual that shares most aspects of a user's self-concept, such as those on matchmaking platforms, the FYP algorithm exposes the user to multiple groups of others who are each similar to the user in at least one specific way. For instance, a user might view content from creators who enjoy baking, other creators who read fantasy books, and still other creators who share the same political views. Each group of content creators is similar to the user in at least one way, for as many facets as they have in their self-concept - a concept we discuss as *diffracted belonging* (see Figure 1). The algorithmic crystal not only reflects multifaceted aspects of one's self-concept back to the user (e.g., a love for baking), but also creates ephemeral connections and facilitates mere belonging to groups of similar others if they share at least one aspect of the self-concept with the user (e.g., other people who love baking), even if they have nothing else in common. Wellman et al. [75] discuss a similar phenomenon of networked individualism where "an individual's overall community will be heterogeneous because people have multiple interests." In the case of our data, however, these connections are not interpersonal relationships; previous work on imitation publics suggests that sociality on social media platforms such as TikTok can be organized around the "shared ritual of content imitation and replication" with strangers, instead of direct interactions with networks of known ties [77]. Similarly, the refractive and diffractive nature of the algorithmic crystal organizes sociality around shared elements of identity, allowing users to surface content from others like them without needing to truly interact with them.

4.2 Applications and Extensions of the Framework

The algorithmic crystal framework serves as an analytic toolkit that provides a grammar for examining how people understand and experience personalized algorithms and a generative framework for future work. Building from an integration of our findings and prior work on algorithms and identity, the framework contributes to the literature by allowing researchers to rethink relevant theories and concepts regarding how people perceive and interact with algorithmic systems. Below, we propose several pathways by which the algorithmic crystal framework can be used to extend theory and inform new lines of research around the implications of algorithms in self-concept development and social life.

First, future work could expand on these ideas by making fine-grained predictions about how people's folk theories of algorithms relate to their behavior. Our data suggests an association between users viewing the algorithm as reflective of their multifaceted and dynamic self-concept, and their enjoyment of the platform. The degree to which algorithms are holistically perceived as crystal-like may thus be related to decisions around continuance of platform use [19]. The individual dimensions of the algorithmic crystal framework can be examined more closely to hone in on how specific perceptions of algorithms (e.g., as more or less multifaceted or dynamic) shape key processes and outcomes, including self-presentation and algorithmic literacy [2, 4, 20]. Future work could also develop ways to assess perceptions of algorithms as crystal-like (e.g., "To what extent does this algorithm reflect your identities and interests as multidimensional?"). Such measurements introduce an opportunity to study between-person differences in perceptions of the same algorithm, between-algorithm differences in user outcomes, and the impact of within-algorithm changes over time - potentially in the context of user adaptation and resistance to algorithmic change (e.g., platform spirit [19]).

The algorithmic crystal can also serve as a useful framework for exploring the ramifications of simplistic algorithmic representations of user identities or those that cannot adapt to an evolving

sense of self. Given the findings of prior work on identity shift and algorithmic mirrors [28], which demonstrate that algorithmically mediated self-presentations can change how people see themselves, the implications of such stereotypical or outdated reflections of the self should be further explored. While misalignment between the actual and algorithmic self may produce dissonance in any individual, reductive reflections of the self (e.g., only being "seen" through one identity, such as gender or race [76]) and outdated reflections of the self (e.g., being "seen" as a role that one may no longer have) may be particularly harmful for certain communities. The failure of an algorithm to capture the self in full as it changes over time may be particularly detrimental for individuals from marginalized identities, for whom self-presentation is of heightened importance [19], and individuals who are undergoing or have experienced a major life change (e.g., adolescents, veterans, widows or divorcees [31]).

On the other hand, the multifaceted and dynamic nature of the crystal may facilitate explorations of how experiences with algorithms may have self-transformative effects. Identity shift refers to the process of self-transformation resulting from intentional self-presentation in a mediated context [10]. While prior research has examined how algorithmic feedback can lead to identity shift by intensifying particular attributes (e.g., introversion, extraversion [49]), the dimensions of the algorithmic crystal may serve as a guide for extant research on more diverse forms of identity shift. For instance, people may experience tension between the algorithmic representation of their true self as produced by organic interactions with the algorithm, and the aspirational version of themselves they may seek to cultivate through strategic refinement. In addition to highlighting distinctions between real and desired selves, engaging reflexively with digital representations of the self has been shown to prompt self-exploration [16] and influence behavior [33]. For instance, students who interacted with altered digital representations of themselves as senior citzens became more conscious of their financial futures and were thus motivated to save more for retirement [60]. What might be the implications of interacting with a multifaceted, dynamic algorithmic representation of the self that can be co-produced and shaped by the user? Applying the framework could allow for nuanced research regarding how personalized content recommendations change people's understanding of who they really are.

Beyond individual identities, Tracy and Trethewey [71] argue that "by conceiving of identities as ongoing, emergent, and not entirely predictable crystals, people are forced to acknowledge a range of possible selves embodied in a range of contexts - even as they are constrained by discourses of power." Similarly, the structural forces embedded in algorithmic systems can also be understood through the lens of the crystal framework. Applying this lens, the algorithmic crystal provides another mechanism that explains why algorithms may overrepresent certain identity facets, while suppressing others. Users may vary in their awareness and perception of this difference, as examined in Karizat et al.'s [38] work on algorithmic representational harms, or the "users experience[s]... of being rendered invisible, trivialized, suppressed, or otherwise further marginalized on the basis of their identities and the algorithm's understanding of their identities." As such, framing the internal algorithmic logics that shape which identities and interests are amplified or minimized through the framework of the crystal could support extant research on biases and harms. For instance, the identity facets that the algorithm presents as the "default" when users first join a platform can reproduce structural norms. In the case of TikTok, the unrefined algorithm promotes content from white, straight, and conventionally attractive youth from the outset, reinforcing the notion that these identities constitute the norm while othering individuals from marginalized identities [38]. Although people can strategically refine the algorithm over time, the burden of engaging in the labor of refinement again disproportionately falls on individuals from marginalized identities.

The algorithmic crystal also contributes a new orientation to understanding social perceptions in algorithmically-structured spaces. Our framework suggests that because personalized algorithms

are organized around an individual's self-concept, some users can come to view others they encounter through the algorithm as refracted images of themselves: an orientation that assumes similarities with visible others based on the folk theory that the algorithm is showing them content reflective of some aspect of their self-concepts. This contributes to our understanding of how people experience different kinds of sociality in algorithmically-curated spaces with specific affordances [2, 6, 77], and demonstrates how folk theorization can shape social experiences. Additional work that characterizes the affect and affiliation dynamics of this orientation to others could contribute meaningfully to our understanding of self-concepts. Further, the framework could be applied to understand how social relationships unfold on platforms affording algorithmic personalization and visually salient content.

The crystal is particularly informative in illustrating the concept of what we call diffracted belonging, a type of algorithmically-mediated connection that facilitates a sense of mere belonging by connecting individuals to n networks of content for n identity facets represented. In the case of TikTok, for example, diffracted belonging appears to be a function of user perceptions that they are iteratively moving from "side to side" and traversing different groups centered around their interests and identities. Diffracted belonging can be understood as a form of social translucence [23], as it provides individuals with partial insight into the ways in which they are connected to other communities of users. In this way, diffracted belonging can be seen as a counterpoint to conceptualizations of algorithmic filter bubbles as stationary, unidimensional echo chambers [13, 47] and a novel extension to psychological theory on mere belonging effects [66]. However, additional work is needed to understand whether such diffractive experiences may recreate dynamics of societal exclusion seen in interpersonal interactions, such as in-group and out-group exclusion [35].

The image of the crystal as being polished or refined by the user connotes agency in one's relationship with the algorithm, and could inspire new design choices. For instance, algorithmically-curated social media content is often described in metaphorical terms like a one-dimensional "stream," where the platform is the active subject that moves while the user stands still. Conversely, consider an image of a person holding a multidimensional crystal that is polished to more accurately reflect the aspects of their self-concept that they want to see, or even rotated to explore new reflections if desired. Rather than positioning users as passive recipients of a feed, the dual processes of algorithmic refinement build upon prior work on strategic curation with algorithms [22, 57] to articulate how individuals can have agency in shaping their algorithmic selves. Further, there are minimal feedback mechanisms for users to reduce unwanted content on their feed, and such mechanisms are typically weak. How might users be able to remove whole networks of content from something like the FYP?

The boundary conditions of the algorithmic crystal framework are likely to be jointly determined through the interaction of both platform affordances and user experience. Indeed, prior work on adaptive folk theorization indicates that human-algorithm engagement is often shaped by structural features, including the technical features (e.g., content modality) of the platform and the psychological characteristics of the user (e.g., algorithmic literacy) [19, 43]. More granular examinations of how networks are organized on particular platforms may prove particularly instructive. Some platforms, for instance, prioritize recommending any content relevant to the user (e.g., TikTok, YouTube), while others prioritize content from known social ties (e.g., Facebook, Instagram). Platforms that more easily afford content-based interactions [42] may better lend themselves to analysis through the algorithmic crystal framework because of the emphasis on personalized content tailored to users' identities. Comparatively, person-based interactions highlight relationships between users and others on the platform and are thus less egocentric. Content

recommendations in these cases may thus be more likely to be understood as derived from shared ties, rather than reflections of the self.

Researchers might also apply the framework by investigating the length and modality of content recommended by the algorithm. For example, an algorithm that hosts shorter videos like TikTok may be more likely to be experienced as multifaceted, because of its ability to reflect a greater number of interest facets in the same length of time. Further, algorithms that prioritize video or image data may be more likely to facilitate feelings of refraction and diffraction because the content provides additional visual and audio cues through which users can view others. Zulli and Zulli [77] describe this through the concept of imitation publics, where videos allow for "physical imitation (copying dance moves), reactive imitation (capitalizing and expanding on someone else's video) and narrative imitation (describing the same type of experiences)."

In terms of user experience, we anticipate that people who have more degrees of algorithmic literacy or algorithmic awareness [19] may be more likely to view algorithms as crystal-like, due to their understanding of how the systems learn from personal engagement to recommend content. Similarly, an individuals' conceptualization of platform spirit, or the underlying motivations of a platform's construction [19], may also be an important factor. A person who views algorithms as exploiting users for monetary gain or manipulating attention may not believe the algorithmically-curated content is an accurate reflection of themselves.

Finally, future work should investigate the contexts in which the algorithmic crystal may "break," or when the framework may not apply. The perception of structural changes in the platform, algorithm, or the companies that produce them may fundamentally shift how individuals understand the algorithm in relation to their self-concept [19]. As prior work on adaptive folk theorization demonstrates, such a shift can be caused by a real-world change (e.g., TikTok updating its content recommendation algorithm) or by a perceived change that is assumed to relate to the algorithm (e.g., a user feeling that the algorithm no longer responds dynamically to their interests). For instance, a change to the TikTok algorithm in late 2021 that prioritized videos with high view-counts caused thousands of users to perceive their algorithm as "broken," after seeing content that they perceived to be more mainstream than their usual fare [30]. On the other hand, users can also perceive their algorithms as "broken" if they feel its recommendations are no longer in alignment with their sense of self. This raises important questions, such as: How much change can an individual accommodate before they no longer experience the algorithm as reflective? Indeed, future work should explore the extent to which crystal-like experiences are temporally-bound, as the product of perceived alignment between two constantly evolving entities: the algorithm and the user.

4.3 Limitations

Limitations of this work include the fact that we only interviewed participants who were students at one of three universities in the Bay Area of California participating in research for course credit. Seeing as location is a factor that influences the content people see on their FYP [69], understanding how participants experience TikTok in various geographic regions, particularly outside the United States, is an important area for future work. We also believe that understanding perceptions of those with marginalized identities, especially those not adequately represented in our sample of participants (e.g., LGBTQ+ participants), will shed additional light on how the algorithmic crystal framework can be applied to issues such as algorithmic representational harms [38].

Further, even though we did not impose any prerequisites regarding how a participant used TikTok - only that they had to be "active" users, our sample included only a few participants who created content as opposed to consuming it. It may be the case that content creators have different perceptions regarding how the algorithm is related to their self-concept, as they are often engaging in acts of self-presentation for an audience [12, 20]. Especially in regard to processes of identity

shift, users interacting with the algorithm through content creation may receive more sources of feedback (e.g., comments or likes on videos) than those who primarily consume content, potentially accentuating self-perception effects.

4.4 Conclusion

We investigated TikTok users' perceptions of the platform, with an emphasis on how people thought about the algorithm in relation to their own self-concept. Through analyzing data from our 24 semi-structured interviews, we propose the *algorithmic crystal framework*. We found that participants generally believed that they were multifaceted and dynamic individuals with interests and identities that were expansive and changed over time. Participants typically felt positively toward the FYP algorithm because the personalized content accurately represented these varied, fluid interests and identities. Further, they believed that their interactions with the platform were influential in aligning their algorithmic selves, or the types of content they saw, with their actual selves. In general, participants also felt a sense of connection towards other users encountered through the algorithm because they were able to recognize parts of themselves refracted in other users and experience ephemeral, diffracted connections with groups of similar others. Lastly, we discussed pathways in which the algorithmic crystal framework can extend theory and inform new lines of research around the implications of algorithms in social life.

REFERENCES

[1] Crystal Abidin. 2021. From "Networked Publics" to "Refracted Publics": A Companion Framework for Researching "Below the Radar" Studies. *Social Media + Society* 7, 1 (Jan. 2021), 205630512098445. https://doi.org/10.1177/2056305120984458

[2] Crystal Abidin. 2021. Mapping Internet Celebrity on TikTok: Exploring Attention Economies and Visibility Labours. *Cultural Science Journal* 12, 1 (Jan. 2021), 77–103. https://doi.org/10.5334/csci.140

[3] Jeffrey Jensen Arnett. 2007. Emerging Adulthood: What Is It, and What Is It Good For? *Child Development Perspectives* 1, 2 (Dec. 2007), 68–73. https://doi.org/10.1111/j.1750-8606.2007.00016.x

[4] Kristen Barta and Nazanin Andalibi. 2021. Constructing Authenticity on TikTok: Social Norms and Social Support on the "Fun" Platform. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–29. https://doi.org/10.1145/3479574

[5] Corey H. Basch, Grace C. Hillyer, and Christie Jaime. 2020. COVID-19 on TikTok: harnessing an emerging social media platform to convey important public health messages. *International Journal of Adolescent Medicine and Health* 0, 0 (Aug. 2020), 20200111. https://doi.org/10.1515/ijamh-2020-0111

[6] Aparajita Bhandari and Sara Bimo. 2022. Why's Everyone on TikTok Now? The Algorithmized Self and the Future of Self-Making on Social Media. *Social Media+ Society* 8, 1 (2022), 20563051221086241.

[7] Sophie Bishop. 2019. Managing visibility on YouTube through algorithmic gossip. *New media & society* 21, 11-12 (2019), 2589–2606.

[8] Matilda Boseley. 2021. 'I found my identity': how TikTok is changing the lives of its popular Indigenous creators. (2021). https://www.theguardian.com/australia-news/2021/jul/10/i-found-my-identity-how-tiktok-is-changing-the-lives-of-its-popular-indigenous-creators

[9] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (Jan. 2017), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

[10] Caleb T Carr, Yeweon Kim, Jacob J Valov, Judith E Rosenbaum, Benjamin K Johnson, Jeffrey T Hancock, and Amy L Gonzales. 2021. An Explication of Identity Shift Theory. *Journal of Media Psychology* (2021).

[11] John Cheney-Lippold. 2017. *We are data: algorithms and the making of our digital selves.* New York University Press, New York.

[12] Sukyoung Choi, Dmitri Williams, and Hyeok Kim. 2020. A snap of your true self: How self-presentation and temporal affordance influence self-concept on social media. *New Media & Society* (Nov. 2020), 146144482097719. https://doi.org/10.1177/1461444820977199

[13] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (March 2021), e2023301118. https://doi.org/10.1073/pnas.2023301118

[14] Charles Horton Cooley. 1902. *Human Nature and the Social Order.* Scribner, New York.

[15] Juliet Corbin and Anselm Strauss. 1998. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

[16] Shelley L Craig and Lauren McInroy. 2014. You can form a part of yourself online: The influence of new media on identity development and coming out for LGBTQ youth. *Journal of Gay & Lesbian Mental Health* 18, 1 (2014), 95–109.

[17] Amber L Cushing. 2011. Self extension and the desire to preserve digital possessions. *Proceedings of the American Society for Information Science and Technology* 48, 1 (2011), 1–3.

[18] Jenny L. Davis. 2020. *How artifacts afford: the power and politics of everyday things*. The MIT Press, Cambridge, Massachusetts.

[19] Michael Ann DeVito. 2021. Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–38. https://doi.org/10.1145/3476080

[20] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173694

[21] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 3163–3174. https://doi.org/10.1145/3025453.3025659

[22] Nicole B Ellison, Penny Trieu, Sarita Schoenebeck, Robin Brewer, and Aarti Israni. 2020. Why We Don't Click: Interrogating the Relationship Between Viewing and Clicking in Social Media Contexts by Exploring the "Non-Click". *Journal of Computer-Mediated Communication* 25, 6 (Dec. 2020), 402–426. https://doi.org/10.1093/jcmc/zmaa013

[23] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.

[24] Moa Eriksson Krutrök. 2021. Algorithmic Closeness in Mourning: Vernaculars of the Hashtag #grief on TikTok. *Social Media + Society* 7, 3 (July 2021), 205630512110423. https://doi.org/10.1177/20563051211042396

[25] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[26] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3174006

[27] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 153–162. https://doi.org/10.1145/2702123.2702556

[28] Megan French. 2018. *Algorithmic Mirrors: an Examination of How Personalized Recommendations Can Shape Self-perceptions and Reinforce Gender Stereotypes*. Ph. D. Dissertation. Stanford University. https://www.proquest.com/dissertations-theses/algorithmic-mirrors-examination-how-personalized/docview/2436884514/se-2?accountid=14026

[29] Anne-Britt Gran, Peter Booth, and Taina Bucher. 2021. To be or not to be algorithm aware: a question of a new digital divide? *Information, Communication & Society* 24, 12 (Sept. 2021), 1779–1796. https://doi.org/10.1080/1369118X.2020.1736124

[30] Palmer Haasch. 2021. TikTok users reported a For You Page glitch that sent them careening into 'straight TikTok'. *Insider* (2021).

[31] Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2016. Digital footprints and changing networks during online identity transitions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2895–2907.

[32] Rebecca A. Hayes, Caleb T. Carr, and Donghee Yvette Wohn. 2016. One Click, Many Meanings: Interpreting Paralinguistic Digital Affordances in Social Media. *Journal of Broadcasting & Electronic Media* 60, 1 (Jan. 2016), 171–187. https://doi.org/10.1080/08838151.2015.1127248

[33] Hal E Hershfield, Daniel G Goldstein, William F Sharpe, Jesse Fox, Leo Yeykelis, Laura L Carstensen, and Jeremy N Bailenson. 2011. Increasing saving behavior through age-progressed renderings of the future self. *Journal of Marketing Research* 48, SPL (2011), S23–S37.

[34] Aaron Hess. 2014. You Are What You Compute (and What is Computed For You): Considerations of Digital Rhetorical Identification. (2014). http://contemporaryrhetoric.com/wp-content/uploads/2017/01/Hess8_1.pdf

[35] Miles Hewstone, Mark Rubin, and Hazel Willis. 2002. Intergroup bias. *Annual review of psychology* 53, 1 (2002), 575–604.

[36] E. Tory Higgins. 1987. Self-discrepancy: A theory relating self and affect. *Psychological Review* 94, 3 (1987), 319–340. https://doi.org/10.1037/0033-295X.94.3.319

[37] Sarah Jerasa and Trevor Boffone. 2021. BookTok 101: TikTok, Digital Literacies, and Out-of-School Reading Practices. *Journal of Adolescent & Adult Literacy* (Oct. 2021), jaal.1199. https://doi.org/10.1002/jaal.1199

[38] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–44. https://doi.org/10.1145/3476046

[39] Becky Kazansky and Stefania Milan. 2021. "Bodies not templates": Contesting dominant algorithmic imaginaries. *new media & society* 23, 2 (2021), 363–381.

[40] Melanie Kennedy. 2020. 'If the rise of the TikTok dance and e-girl aesthetic has taught us anything, it's that teenage girls rule the internet right now': TikTok celebrity, girls and the Coronavirus crisis. *European Journal of Cultural Studies* 23, 6 (Dec. 2020), 1069–1076. https://doi.org/10.1177/1367549420945341

[41] Vanessa Kitzie. 2019. "That looks like me or something i can do": Affordances and constraints in the online identity work of US LGBTQ+ millennials. *Journal of the Association for Information Science and Technology* 70, 12 (Dec. 2019), 1340–1351. https://doi.org/10.1002/asi.24217

[42] Daniel Klug, Yiluo Qin, Morgan Evans, and Geoff Kaufman. 2021. Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm. In *13th ACM Web Science Conference 2021*. ACM, Virtual Event United Kingdom, 84–92. https://doi.org/10.1145/3447535.3462512

[43] Abby Koenig. 2020. The algorithms know me and i know them: using student journals to uncover algorithmic literacy awareness. *Computers and Composition* 58 (2020), 102611.

[44] Mark R. Leary. 1996. *Self-presentation: impression management and interpersonal behavior*. Westview Press, Boulder, Colo.

[45] Amalie MacGowan. 2020. The TikTok Algorithm Knew My Sexuality Better Than I Did. https://repeller.com/tiktok-algorithm-bisexual/

[46] Matthew B. Miles, A. M. Huberman, and Johnny Saldaña. 2020. *Qualitative data analysis: a methods sourcebook* (fourth edition ed.). SAGE, Los Angeles.

[47] Eli Pariser. 2011. *The filter bubble: what the Internet is hiding from you*. Penguin Press, New York. OCLC: ocn682892628.

[48] Chang Sup Park and Barbara K Kaye. 2019. Smartphone and self-extension: Functionally, anthropomorphically, and ontologically extending self via the smartphone. *Mobile Media & Communication* 7, 2 (2019), 215–231.

[49] Gwen Petro. 2021. *Adding Insult to Algorithm: How Unfavorable Behavioral Advertising Impacts Self-Concept*. University of California, Santa Barbara.

[50] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul, Republic of Korea, 173–182. https://doi.org/10.1145/2702123.2702174

[51] Morgan Quinn Ross and Joseph B Bayer. 2021. Explicating self-phones: Dimensions and correlates of smartphone self-extension. *Mobile Media & Communication* (2021), 2050157920980508.

[52] Andreas Schellewald. 2022. Theorizing "Stories About Algorithms" as a Mechanism in the Formation and Maintenance of Algorithmic Imaginaries. *Social Media+ Society* 8, 1 (2022), 20563051221077025.

[53] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big data & society* 4, 2 (2017), 2053951717738104.

[54] Nick Seaver. 2019. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture* 24, 4 (Dec. 2019), 421–436. https://doi.org/10.1177/1359183518820366

[55] Donghee Shin, Bu Zhong, and Frank A. Biocca. 2020. Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management* 52 (June 2020), 102061. https://doi.org/10.1016/j.ijinfomgt.2019.102061

[56] Ignacio Siles, Andrés Segura-Castillo, Ricardo Solís, and Mónica Sancho. 2020. Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society* 7, 1 (2020), 2053951720923377.

[57] Ignacio Siles González and Ariana Meléndez Moran. 2021. "The most aggressive of algorithms": User awareness of and attachment to TikTok's content personalization. https://hdl.handle.net/10669/83230

[58] Ellen Simpson, Andrew Hamann, and Bryan Semaan. 2022. How to Tame: LGBTQ+ Users' Domestication of TikTok. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–27.

[59] Ellen Simpson and Bryan Semaan. 2021. For You, or For"You"?: Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 1–34. https://doi.org/10.1145/3432951

[60] Tamara Sims, Sarah Raposo, Jeremy N Bailenson, and Laura L Carstensen. 2020. The future is now: Age-progressed images motivate community college students to prepare for their financial futures. *Journal of Experimental Psychology: Applied* 26, 4 (2020), 593.

[61] Ben Smith. 2021. How TikTok Reads Your Mind. (2021). https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html

[62] Oren Soffer. 2021. Algorithmic Personalization and the Two-Step Flow of Communication. *Communication Theory* 31, 3 (Sept. 2021), 297–315. https://doi.org/10.1093/ct/qtz008

[63] WSJ Staff. 2021. Inside TikTok's Algorithm: A WSJ Video Investigation. (2021). https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477

[64] Nicole Strimbu and Michael O'Connell. 2019. The Relationship Between Self-Concept and Online Self-Presentation in Adults. *Cyberpsychology, Behavior, and Social Networking* 22, 12 (Dec. 2019), 804–807. https://doi.org/10.1089/cyber.2019.0328

[65] Yiran Su, Bradley J. Baker, Jason P. Doyle, and Meimei Yan. 2020. Fan Engagement in 15 Seconds: Athletes' Relationship Marketing During a Pandemic via TikTok. *International Journal of Sport Communication* 13, 3 (Sept. 2020), 436–446. https://doi.org/10.1123/ijsc.2020-0238

[66] Henri Tajfel and John C. Turner. 1986. The social identity theory of intergroup behavior. In S. Worchel and WG Austing (Eds.). Psychology of intergroup relations. (1986).

[67] Leila Takayama. 2009. Making sense of agentic objects and teleoperation: In-the-moment and reflective perspectives. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 239–240.

[68] Kaitlyn Tiffany. 2021. I'm Scared of the Person TikTok Thinks I Am. (2021). https://www.theatlantic.com/technology/archive/2021/06/your-tiktok-feed-embarrassing/619257/

[69] TikTok. 2020. How TikTok recommends videos #ForYou. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

[70] Sarah J. Tracy and Sophia Town. 2020. Real, Fake, and Crystallized Identities. In *The Oxford Handbook of Identities in Organizations*, Andrew D. Brown (Ed.). Oxford University Press, 390–407. https://doi.org/10.1093/oxfordhb/9780198827115.013.3

[71] Sarah J. Tracy and Angela Trethewey. 2005. Fracturing the Real-Self?Fake-Self Dichotomy: Moving Toward "Crystallized" Organizational Discourses and Identities. *Communication Theory* 15, 2 (May 2005), 168–195. https://doi.org/10.1111/j.1468-2885.2005.tb00331.x

[72] Laura Vandenbosch and Steven Eggermont. 2016. The Interrelated Roles of Mass Media and Social Media in Adolescents' Development of an Objectified Self-Concept: A Longitudinal Study. *Communication Research* 43, 8 (Dec. 2016), 1116–1140. https://doi.org/10.1177/0093650215600488

[73] Maike Vollstedt and Sebastian Rezat. 2019. An introduction to grounded theory with a special focus on axial coding and the coding paradigm. *Compendium for early career researchers in mathematics education* 13 (2019), 81–100.

[74] Gregory M. Walton, Geoffrey L. Cohen, David Cwir, and Steven J. Spencer. 2012. Mere belonging: The power of social connections. *Journal of Personality and Social Psychology* 102, 3 (2012), 513–532. https://doi.org/10.1037/a0025731

[75] Barry Wellman, Anabel Quan-Haase, Jeffrey Boase, Wenhong Chen, Keith Hampton, Isabel Díaz, and Kakuko Miyata. 2003. The Social Affordances of the Internet for Networked Individualism. *Journal of Computer-Mediated Communication* 8, 3 (April 2003), 0–0. https://doi.org/10.1111/j.1083-6101.2003.tb00216.x

[76] Brita Ytre-Arne and Hallvard Moe. 2021. Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society* 43, 5 (2021), 807–824.

[77] Diana Zulli and David James Zulli. 2020. Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media & Society* (Dec. 2020), 146144482098360. https://doi.org/10.1177/1461444820983603

PX0320

Meta

Get the from Instagram

Start getting inspired with announcements, tips, and success stories on our blog.

TIP NOVEMBER 1, 2016

BY: Instagram Business Team
San Francisco, CA

Reaching Your Customers on Instagram

*(Update on November 14, 2021 at 8:30AM PT: This blog post was published in 2016. For up-to-date information about reaching your audience on Instagram, learn more about **ads on Instagram**.)*

Whether your audience is ice cream lovers in Southern California, cattle ranchers in South America, or car

enthusiasts around the world, audience targeting helps you show your ads to the people you care about most.

Facebook Page, you have the opportunity to target your ads to a specific audience. This audience can be based on location and demographics like age, gender and interests. You can even target your ad to people based on what they do off of Instagram.

It's up to you to choose the audience you want to reach. You can choose from one, or a combination of targeting options that suit your business' needs. By targeting your ads, people receiving the ads are ones more likely to be interested in your product, app or service.

In addition to targeting ads my location and demographic, you have more targeting options based on connections.

- **Custom Audiences**: Create an audience of your existing customers.

- **Lookalike Audiences**: Reach a new audience of people who are similar to customers you care about.

- Automatic Targeting: We help you quickly create an audience who might be interested in your business using a variety of signals including location, demographics and interests.

With a half a billion users on Instagram from around the world, targeting your ad helps you reach the ones who matter to you most

Prev article Next article

Goals

Products

Inspiration

Other Services

English (US) English (UK) Español Português (Brasil)

Français (France) Español (España) العربية Tiếng Việt Italiano

Deutsch ภาษาไทย 中文(台灣) Čeština Dansk Ελληνικά

Suomi Français (Canada) עברית हिन्दी Magyar

Bahasa Indonesia 日本語 한국어 Bahasa Melayu

Privacy Terms Help

© 2024 Meta

PX0321

TikTok is discontinuing its BeReal clone

TikTok is discontinuing its BeReal clone

/

The change comes as viral sensation BeReal fades from the spotlight.

By Jon Porter, a reporter with five years of experience covering consumer tech releases, EU tech policy, online platforms, and mechanical keyboards.
Jun 27, 2023, 6:47 AM EDT

Share this story

-
-
-

Illustration by Nick Barclay / The Verge

TikTok is killing off its BeReal clone TikTok Now, according to notifications being sent to users. Multiple Twitter users have posted screenshots of the message in different languages, which says that ByteDance is "updating the TikTok experience and [is] discontinuing TikTok Now." News of its discontinuation comes a little over nine months after TikTok Now was officially announced, although it's unclear exactly when the feature will disappear for good.

TikTok Now was launched in September with the stated aim of fostering "authentic and spontaneous connections on TikTok." But it was hard to ignore the fact that its format (asking users to capture a moment once a day using the front and back cameras of their phone) was all but identical to BeReal, which exploded in popularity last year.

TikTok's twist was that TikTok Now supported up to 10-second-long videos in addition to just still photos. In the US, TikTok Now has been available in the main TikTok app, while owner ByteDance also launched a dedicated TikTok Now app in other regions around the world. Social media consultant Matt Navarra confirmed to *The Verge* that he was shown the notification announcing the discontinuing of TikTok Now within the main TikTok app on iOS.

Although TikTok's notification doesn't give an exact reason for why TikTok Now is being discontinued, it's hard to ignore the fact that BeReal isn't the viral sensation it once was. "They're over being real," was *The New York Times'* conclusion in April, when it cited a 61 percent drop in BeReal daily active users between October 2022 and March 2023 according to third-party data. Per Apptopia, user numbers had dropped from around 15 million to under six million people in March, the *NYT* reported.

Although BeReal disputes these numbers ("BeReal has over 20 million daily active users around the world" it wrote in April), anecdotally I'm seeing far fewer mentions of the app on social media platforms these days. Regardless of the exact figures, it feels like the viral spotlight has moved on.

TikTok hasn't been the only social media company to experiment with adding BeReal-style features to its app. SnapChat launched a dual-camera feature in August, which traded the daily notification and limits of BeReal for more editing and layout options. Meanwhile Instagram has also been testing its own take on the feature called "Candid Stories." My colleague Casey Newton has taken to referring to these kinds of copycat features as "murder clones." In other words, they represent an attempt by dominant social media apps to fend off would-be competitors by aping their unique features.

A spokesperson for TikTok did not immediately respond to *The Verge*'s request for comment. As of this writing, the service's support page about TikTok Now has not been updated to mention a shutdown.

PX0322

PLOS ONE

RESEARCH ARTICLE

Don't put all social network sites in one basket: Facebook, Instagram, Twitter, TikTok, and their relations with well-being during the COVID-19 pandemic

Alexandra Masciantonio[1]*, **David Bourguignon**[1], **Pierre Bouchat**[1], **Manon Balty**[1], **Bernard Rimé**[2]

1 Department of Psychology, Université de Lorraine, Metz, France, 2 Department of Psychology, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

* Alexandra.masciantonio@univ-lorraine.fr

Abstract

Prior studies indicated that actively using social network sites (SNSs) is positively associated with well-being by enhancing social support and feelings of connectedness. Conversely, passively using SNSs is negatively associated with well-being by fostering upward social comparison and envy. However, the majority of these studies has focused on Facebook. The present research examined the relationships between well-being—satisfaction with life, negative affect, positive affect—and using actively or passively various SNSs—Facebook, Instagram, Twitter, TikTok—during the COVID-19 pandemic. In addition, two mediators were tested: social support and upward social comparison. One thousand four persons completed an online survey during the quarantine measures; the analyses employed structural equation modeling. Results showed that passive usage of Facebook is negatively related to well-being through upward social comparison, whereas active usage of Instagram is positively related to satisfaction with life and negative affect through social support. Furthermore, active usage of Twitter was positively related to satisfaction with life through social support; while passive usage was negatively related to upward social comparison, which, in turn, was associated with more negative affect. Finally, TikTok use was not associated with well-being. Results are discussed in line with SNSs' architectures and users' motivations. Future research is required to go beyond methodological and statistical limitations and allow generalization. This study concludes that SNSs must be differentiated to truly understand how they shape human interactions.

OPEN ACCESS

Citation: Masciantonio A, Bourguignon D, Bouchat P, Balty M, Rimé B (2021) Don't put all social network sites in one basket: Facebook, Instagram, Twitter, TikTok, and their relations with well-being during the COVID-19 pandemic. PLoS ONE 16(3): e0248384. https://doi.org/10.1371/journal.pone.0248384

Editor: Barbara Guidi, University of Pisa, ITALY

Received: September 4, 2020

Accepted: February 26, 2021

Published: March 11, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0248384

Copyright: © 2021 Masciantonio et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available in OSF (Open Science Framework) at: https://doi.org/10.17605/OSF.IO/S5MJX.

Introduction

The COVID-19 pandemic that has hit the world since the end of 2019 has led the governments of many countries to impose quarantine measures on their populations. For many people, these confinement measures led to a drastic reduction in interpersonal relations. However,

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

interpersonal relations have powerful beneficial effects on physical and mental health [1,2]. In order to cope with the negative effects of social isolation on well-being, a significant number of recommendations were issued [3,4]. Several of them, derived predominantly from non-scholars, have promoted the use of social network sites (SNSs) to keep contact with family and friends [5,6]. Nevertheless, this assertion involves addressing one complex question: are SNSs really beneficial to well-being? The present study will contribute to this research question, through a short literature review and an empirical study. More research will be required to provide a clear understanding of how SNSs impact well-being.

Definition of key concepts

Before addressing the relationship between SNSs and well-being, both need to be defined. On one side, the familiar definition of Ellison and boyd [7 p157] described SNSs as networked communication platforms in which participants 1) have uniquely identifiable profiles that consist of user-supplied content, content provided by other users, and/or system-level data; 2) can publicly articulate connections that can be viewed and traversed by others; and 3) can consume, produce, and/or interact with streams of user-generated content provided by their connections on the site.

On the other side, the term well-being refers, in this study, to the subjective part of well-being. Instead of relying on physical and material resources, subjective well-being can be understood as "people's overall evaluations of their lives and their emotional experiences." [8 p87]. Hence, subjective well-being is a multidimensionality construct, and each component needs to be assessed individually. Typically, subjective well-being comprises at least three components: a sense of satisfaction with life, the presence of positive affect, and the absence of negative affect [8]. The satisfaction with life allows to capture how people evaluate their lives (i.e., cognitive level of subjective well-being). Likewise, positive and negative affect reflect what feelings people experience in their lives (i.e., affective level of subjective well-being).

Literature review

The relation between SNSs and well-being may at first seem inconsistent. Several studies showed that SNSs use is negatively associated with well-being [9,10], while others revealed a positive relationship [11,12]. However, these studies relied on an overall measure of SNSs use, whereas two distinguish usages can be proposed: an active (e.g., interacting directly with others by posting content or commenting others' content) and a passive one (e.g., reading and consuming others' content). Gerson, Plagnol and Corr [13] pointed out how these usages match with specific SNS activities, demonstrating they reflect related, but separate constructs. In that respect, results seem more uniform when different modalities of SNSs use have been taken into account: actively using SNSs is positively associated with well-being and in contrast, passively using SNSs is negatively associated with well-being [14–19]. Verduyn, Ybarra, Résibois, Jonides and Kross [20] reviewed the literature and identified the mechanisms underlying these relationships. Their model suggests that actively using SNSs increases subjective well-being by improving social capital and feelings of connectedness. Conversely, passively using SNSs lessens subjective well-being by fostering social comparison and envy. Although this model is a major step to clarify the consequences of SNSs on well-being, most of the studies underpinning these mechanisms have focused on Facebook. Doing so, one can wonder whether other SNSs might have different impacts on well-being.

Few studies have investigated the effects of different SNSs on well-being. Pittman and Reich [21] demonstrated that the use of image-based platforms (e.g., Instagram, Snapchat) was positively associated with well-being and negatively with loneliness, whereas text-based platforms

(e.g., Twitter, Yik Yak) were not related to well-being and loneliness. Recently, Chae [22] has also examined the relationships between various platforms and well-being through social comparison. As expected, social comparison was negatively associated with well-being; but while Instagram and LinkedIn enhanced social comparison, Twitter decreased it. Surprisingly, Facebook use was not related to well-being. These two studies have therefore yielded contradictory outcomes, but they employed an overall measure of SNSs use which makes impossible to investigate the distinct effects of passive and active usages.

Overview of the research

To draw conclusions on SNSs and well-being, the literature on passive and active usage need to be integrated with the literature on cross-media studies. To that end, the present research examines the relationships between various SNSs and well-being through two mediators—social support and upward social comparison. Specifically, this study focuses on the active and passive usages of four popular SNSs: Facebook, Instagram, Twitter and TikTok [23]. Although Facebook, Instagram and Twitter are henceforth well studied in the literature, TikTok is a new SNS created in 2016 with a number of users increasing day by day [24]. These SNSs differ from each other by their architectures [25]: Facebook incorporates both image and text, Twitter is text-based, and Instagram as well as TikTok are image-based (the first concerns pictures and the second relies on videos). Instagram, TikTok and Twitter are also unidirectional (i.e., possibility to follow someone's content without their approval), whereas Facebook is dyadic (i.e., need to be approved by someone to access their content). Moreover, people do not use them for the same reasons: Facebook use is mainly related to social support and self-presentation [26]; Instagram allows users to self-document, self-promote, express one's creativity and see other's content [27]; Twitter use is mainly driven by informational needs [28,29]. Finally, only one study examined TikTok use and concluded that the platform was seen as a "recording tool rather than a social media app" [30 p132]. Indeed, self-document was the most important motivation to use TikTok.

The model of Verduyn et al. [20] is mainly based on Facebook use, one would therefore expect to draw the same conclusions as the authors:

> Hypothesis 1: Social support mediates the positive association between actively using Facebook and subjective well-being, and upward social comparison mediates the negative association between passively using Facebook and subjective well-being.

Image-based SNSs, such as Instagram and TikTok, have been shown to be related to well-being [21]. Moreover, Instagram users want to keep in touch with their friends, but also to self-promote [27]. Hence, social support and upward social comparison could both play a part in this relation. One would therefore expect the model of Verduyn et al. [20] to be generalized to Instagram:

> Hypothesis 2: Social support mediates the positive association between actively using Instagram and subjective well-being, and upward social comparison mediates the negative association between passively using Instagram and subjective well-being.

In contrast, TikTok use was not firstly motivated by social interaction or self-presentation [30]. So, no assumption can be made about the mediating roles of social support and upward social comparison. The only hypothesis which can be proposed is:

Hypothesis 3: Actively using TikTok is positively associated with well-being and passively using TikTok is negatively associated with well-being.

Finally, text-based SNSs do not appear to be related to well-being [21]. The following hypothesis is therefore proposed for Twitter:

Hypothesis 4: Actively and passively using Twitter is not associated with well-being.

Method

Participants and procedure

One thousand four persons agreed to participate in the study. Among them, were excluded those reporting missing data and under the age of 18. The final sample was composed of 793 participants (613 women, 178 men and 2 persons who have a gender identity other than male or female) aged between 18 and 77 years old (M = 33.75, SD = 14.70). All participants were francophone: 463 were French, 264 were Belgian, 20 were Swiss and 46 had another nationality. Regarding the highest degree completed, one person had no primary education, 253 had a high school degree, 285 had a university short cycle degree (two or three years), 207 had a university long cycle degree (four or five years) and 47 had a doctorate. Finally, 89% had a Facebook account (N = 703), 63% had an Instagram account (N = 502), 38% had a Twitter account (N = 300) and 15% had a TikTok account (N = 121). An anonymous online survey was created using the Qualtrics Survey Software. Participants were recruited through academic mailing lists from social science, which explains the large proportion of women and academic people in the sample. Before completing the measures, all participants signed an informed consent form and accepted voluntary to take part in this research. Data collection was carried out from 7th April 2020 to 16th April 2020. Measures reported in the present study are part of a larger questionnaire; all data are available in OSF (Open Science Framework) at: https://osf.io/s5mjx/.

Measures

Overall SNS use: When participants declared to have an account for one of the four SNSs (Facebook, Instagram, Twitter, TikTok), they indicated the frequency they used this SNS before and during the quarantine measures on a 7-point scale (never; between one and three times a year; less than once a month; one to four times a month; one to four times a week; one to three times a day; more than three times a day).

Active and passive usage of SNSs: To be consistent with the literature [13], we chose to measure passive and active usage as separate constructs. This means that users can have both an active and a passive SNS usage; they can spend most of their time scrolling their news feed, but they can also send messages throughout the day. When participants declared to have an account for a SNS (Facebook, Instagram, Twitter, TikTok), they were therefore asked how much they used this SNS actively (1 = not actively at all; 7 = very actively) and passively (1 = not passively at all; 7 = very passively) during the quarantine measures [19]. Active usage was defined as "posting and commenting on [Facebook][Instagram][Twitter][TikTok], for example: post content on your profile, react to posts and comments from other users, etc.", while passive usage as "browsing [Facebook][Instagram][Twitter][TikTok], for example: scrolling through your news feed, looking at other users' profiles, etc.".

Motivations to use SNSs: Three motivations to use SNSs were derived from Cheung, Chiu and Lee [31]: maintaining interpersonal interconnectivity ("To stay in touch"; "To have something to do with others"), purposive value ("To get information"; "To provide others with

information") and entertainment value ("To pass time away when bored"; "To be entertained"). Participants were asked to rate the extent to which these 6 items correspond to their motivations to use SNSs during the quarantine measures on a 7-point Likert scale (1 = strongly disagree; 7 = strongly agree). McDonald's ω computed a value of.80 for maintaining interpersonal interconnectivity,.57 for purposive value and.79 for entertainment value.

Social support on SNSs: Social support on SNSs was measured using eight items adapted from Nick et al. [32]. Two items were chosen for each subscale (emotional support, informational support, social companionship and instrumental support). Participants indicated their agreement with these items on a 7-point Likert scale. Sample items include "During quarantine measures, people show that they care about me on social network sites." and "During quarantine measures, people give me useful advice on social network sites.". Scores for each subscale were averaged such that a higher overall score indicated greater social support on SNSs during the quarantine measures (McDonald's ω = .83).

Upward social comparison: The upward social comparison was inspired from Brunot and Juhel [33] and consisted in two items: "On social network sites, I sometimes think that my relatives (friends, family and colleagues) are fare better than me during the quarantine measures" and "On social network sites, I sometimes think that my relatives (friends, family and colleagues) are better off than me". Participants indicated their agreement with these items on a 7-point Likert scale (McDonald's ω = .84).

Positive affect: Positive affect were assessed by asking participants to rate of much they feel "Optimistic, encouraged, hopeful" and "Proud, trustful, self-confident" on a 7-point Likert scale (McDonald's ω = .76). The measure was adapted from Fredrickson [34].

Negative affect: Negative affect were assessed by asking participants to rate of much they feel "Sad, depressed, unhappy", "Angry, furious" and "Anxious, frightened" on a 7-point Likert scale (McDonald's ω = .75). The measure was adapted from Gaudreau, Sanchez and Blondin [35].

Satisfaction with life: Satisfaction with life was measured using the Satisfaction with Life Scale [36]. An example item is "I am satisfied with my life". Participants indicated their agreement with the five items on a 7-point Likert scale. Given the good reliability (McDonald's ω = .89), the five items were aggregated.

Results

Analyses were conducted using the JASP software [37].

Preliminary analyses

An exploratory analysis of the data is available in OSF at: https://osf.io/fe4pn/. Four paired samples T-Tests have been also carried out between the overall SNS use before the quarantine measures and during the quarantine measures. Results showed that the overall use have increased during the quarantine for all SNSs, and in particular for TikTok: Facebook ($t_{(702)}$ = 11.84, $p < .001$, d = .45), Instagram ($t_{(501)}$ = 6.33, $p < .001$, d = .28), Twitter ($t_{(299)}$ = 4.02, $p < .001$, d = .23) and TikTok ($t_{(120)}$ = 10.31, $p < .001$, d = .94). Finally, correlations between overall SNS use during quarantine and motivations to use SNSs are presented in Table 1.

Main analyses

Structural equation modeling with Lavaan [38] was used to examine the relationships between SNSs (Facebook, Instagram, Twitter, TikTok) and well-being (positive affect, negative affect and satisfaction with life) through two mediators, social support and upward social comparison. For each model, dependent variables were controlled for age and gender. A one step

Table 1. Pearson's correlations between overall SNSs use during the quarantine measures and motivations to use SNSs.

	Maintaining interpersonal interconnectivity	Purposive value	Entertainment value
Overall **Facebook** use during quarantine	**.101****	**.076***	**.135*****
Overall **Instagram** use during quarantine	**.100***	.062	**.376*****
Overall **Twitter** use during quarantine	.008	**.147***	**.307*****
Overall **TikTok** use during quarantine	.132	.090	**.229***

Note.
*$p < .05$;
**$p < .01$;
***$p < .001$.

https://doi.org/10.1371/journal.pone.0248384.t001

approach was employed, that means that the parameters of the measurement model and the structural model were estimated simultaneously. Analyses were carried on with DWLS (diagonally weighted least squares) estimator which is adapted for data violating normality [39]. Five fit indices were chosen: χ^2 (chi-square), SRMR (Standard Root Mean Square Residuals), RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index) and TLI (Tucker-Lewis Index) [39]. The first two address the global fit of the model: χ^2 must be nonsignificant and the value of SRMR must be equal or lower to.08. RMSEA concerns the parsimony of the model and must be lower to.06. Lastly, CFI and TLI are incremental indices and must be superior to.9.

Facebook. All standardized item loadings exceeded.4 and were significant ($p < .001$). The results also revealed a satisfactory model fit to the data: $\chi^2(227, N = 703) = 723.86$, $p < .001$; SRMR = .06; RMSEA = .056; CFI = .948; TLI = .938. Although the χ^2 is significant, this statistic is very sensitive to sample size [39].

As shown in Fig 1, direct paths from actively and passively using Facebook to satisfaction with life, positive affect and negative affect were nonsignificant ($p > .05$), except the path from using actively Facebook to negative affect ($\beta = .17$, $p < .05$). Contrary to hypothesis 1, direct path from actively using Facebook to social support was nonsignificant ($p > .05$), but direct path from passively using Facebook to upward social comparison was significant ($\beta = .13$, $p < .05$). All estimated paths from social support and upward social comparison to the three constructs of well-being were significant ($p < .05$). Consistent with hypothesis 1, the indirect effects of passively using Facebook on well-being (satisfaction with life, positive affect and negative affect) through upward social comparison were significant ($p < .05$).

In other words, results revealed that upward social comparison mediates the negative association between passively using Facebook and subjective well-being. Nonetheless, using actively Facebook was also directly associated with greater negative affect.

Instagram. All standardized items loadings exceeded.4 and were significant ($p < .001$). The results also revealed a satisfactory model fit to the data: $\chi^2(227, N = 502) = 532.32$, $p < .001$; SRMR = .061; RMSEA = .052; CFI = .955; TLI = .947.

As shown in Fig 2, direct paths from actively and passively using Instagram to satisfaction with life, positive affect and negative affect were nonsignificant ($p > .05$). Contrary to hypothesis 2, direct path from passively using Instagram to upward social comparison was nonsignificant ($p > .05$), but direct path from actively using Instagram to social support was significant ($\beta = .21$, $p < .05$). All estimated paths from social support and social comparison to the three constructs of well-being were significant ($p < .05$), except the path from social support to positive affect which was nonsignificant ($p > .05$). Partially consistent with hypothesis 2, the

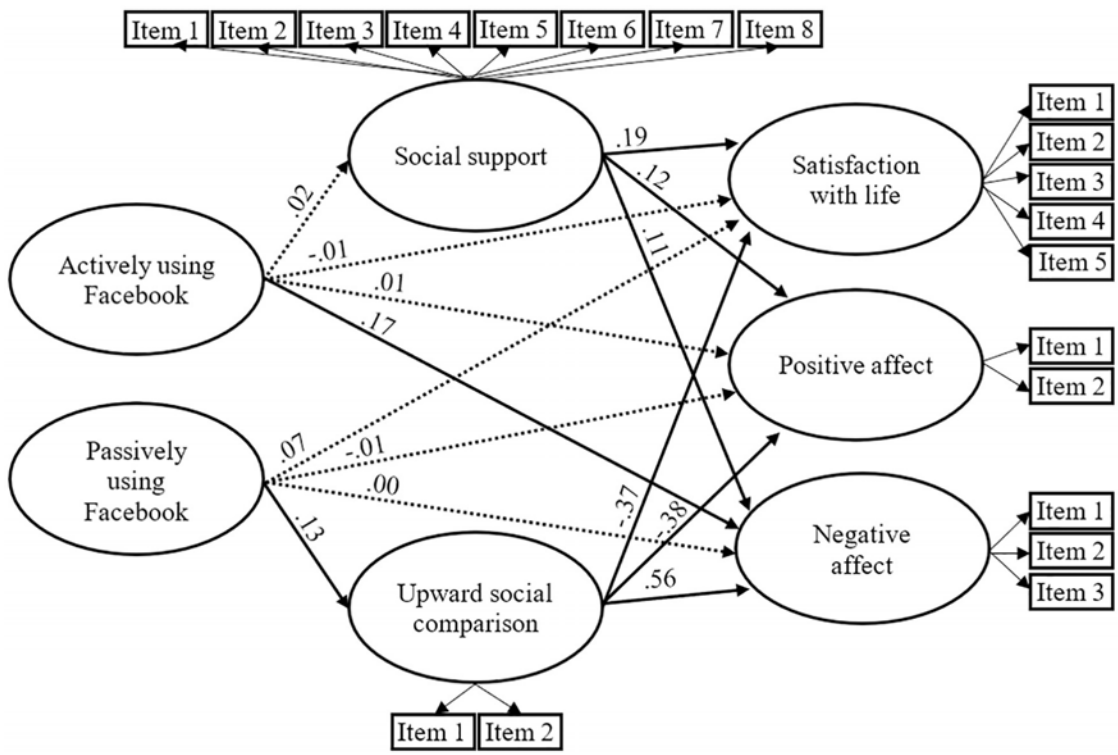


Fig 1. The estimated standardized parameters of the Facebook model. Dashed lined indicate nonsignificant paths ($p > .05$). The three components of well-being were controlled–but not displayed—for gender and age: Age was associated with satisfaction with life (β = .26, $p < .05$), negative affect (β = -.33, $p < .05$), and positive affect (β = .20, $p < .05$); women had less positive affect (β = -.18, $p < .05$), and more negative affect (β = .12, $p < .05$).

https://doi.org/10.1371/journal.pone.0248384.g001

indirect effects of actively using Instagram on satisfaction with life and negative affect through social support was significant ($p < .05$).

In other words, results revealed that social support mediates the positive association between actively using Instagram and satisfaction with life on one hand, and the positive association between actively using Instagram and negative affect on the other.

Twitter. All standardized items loadings exceeded.4 and were significant ($p < .001$). The results also revealed a satisfactory model fit to the data: $\chi^2(227, N = 300) = 415.61$, $p < .001$; SRMR = .071; RMSEA = .053; CFI = .953; TLI = .944.

As shown in Fig 3, direct paths from actively and passively using Twitter to satisfaction with life, positive affect and negative affect were nonsignificant ($p > .05$). Contrary to hypothesis 4, direct path from passively using Twitter to upward social comparison was significant (β = -.14, $p < .05$), and direct path from actively using Twitter to social support was also significant (β = .15, $p < .05$). All estimated paths from social support and social comparison to the three constructs of well-being were significant ($p < .05$), except the path from social support to positive affect which was nonsignificant ($p > .05$). The indirect effect of actively using Twitter on satisfaction with life through social support was significant ($p < .05$). Likewise, the indirect effects of passively using Twitter on negative affect through upward social comparison was significant ($p < .05$), and the indirect effects of passively using Twitter on satisfaction with life and positive affect through upward social comparison was significant were near significant ($p = .06$ for satisfaction with life; $p = .056$ for positive affect).

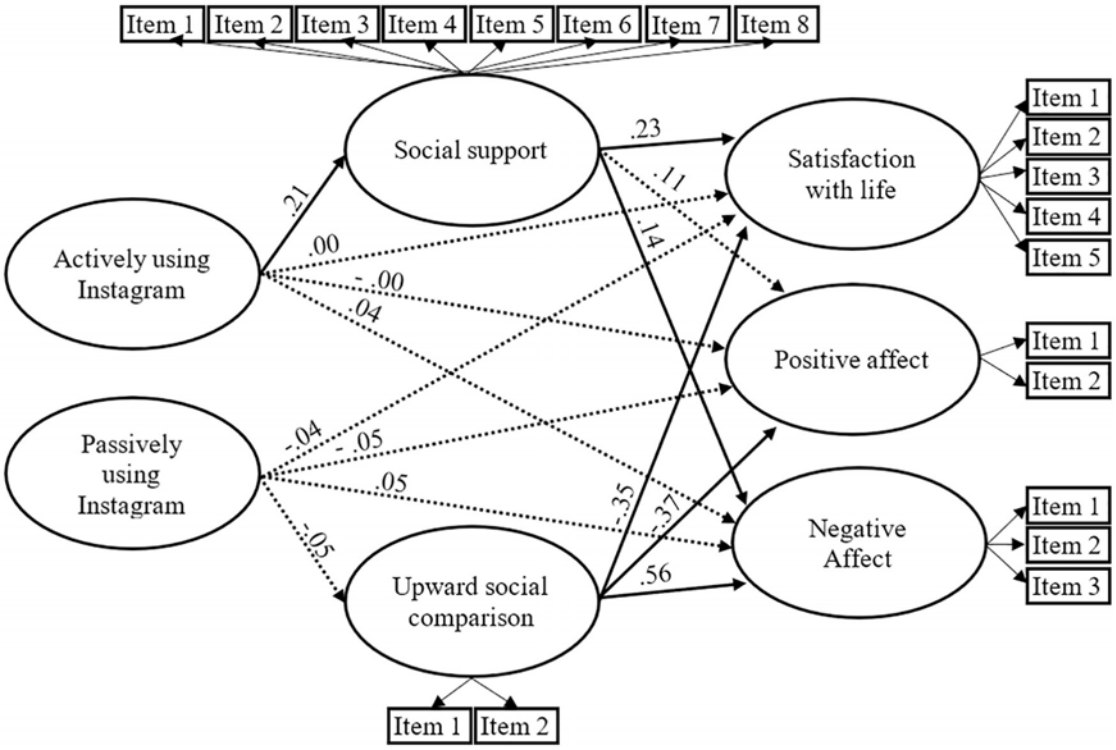


Fig 2. The estimated standardized parameters of the Instagram model. Dashed lined indicate nonsignificant paths ($p > .05$). The three components of well-being were controlled–but not displayed—for gender and age: Age was associated with satisfaction with life (β = .27, $p < .05$), negative affect (β = -.28, $p < .05$), and positive affect (β = .27, $p < .05$); women had less positive affect (β = -.18, $p < .05$), and more negative affect (β = .11, $p < .05$).

https://doi.org/10.1371/journal.pone.0248384.g002

In other words, results showed that actively using Twitter was associated with more social support, and that using passively Twitter was associated with less upward social comparison. In addition, social support mediated the relation between using actively Twitter and satisfaction with life, and social comparison mediated the relation between passively using Twitter and negative affect.

TikTok. All standardized items loadings were significant ($p < .05$) and exceeded.4, except one item of the social support construct (.24). The results also revealed that the model fits the data well: $\chi^2(227, N = 121) = 227.034$, $p > .05$; SRMR = .084; RMSEA = .001; CFI = 1.000; TLI = 1.000.

As shown in Fig 4 and inconsistent with hypothesis 3, direct paths from actively and passively using TikTok to satisfaction with life, positive affect and negative affect were nonsignificant ($p > .05$). Besides, direct paths from passively and actively using Twitter to upward social comparison and social support respectively, were nonsignificant ($p > .05$). Lastly, only paths from upward social comparison to positive and negative affect were significant ($p < .05$).

In other words, results revealed that actively and passively using Tiktok was not associated with well-being, and that social support and upward social comparison did not appear to play a meditational role between TikTok use and well-being.

Discussion

Past researches have shown that actively using SNSs is positively associated with well-being through social support, and that passively using SNSs is negatively associated with well-being

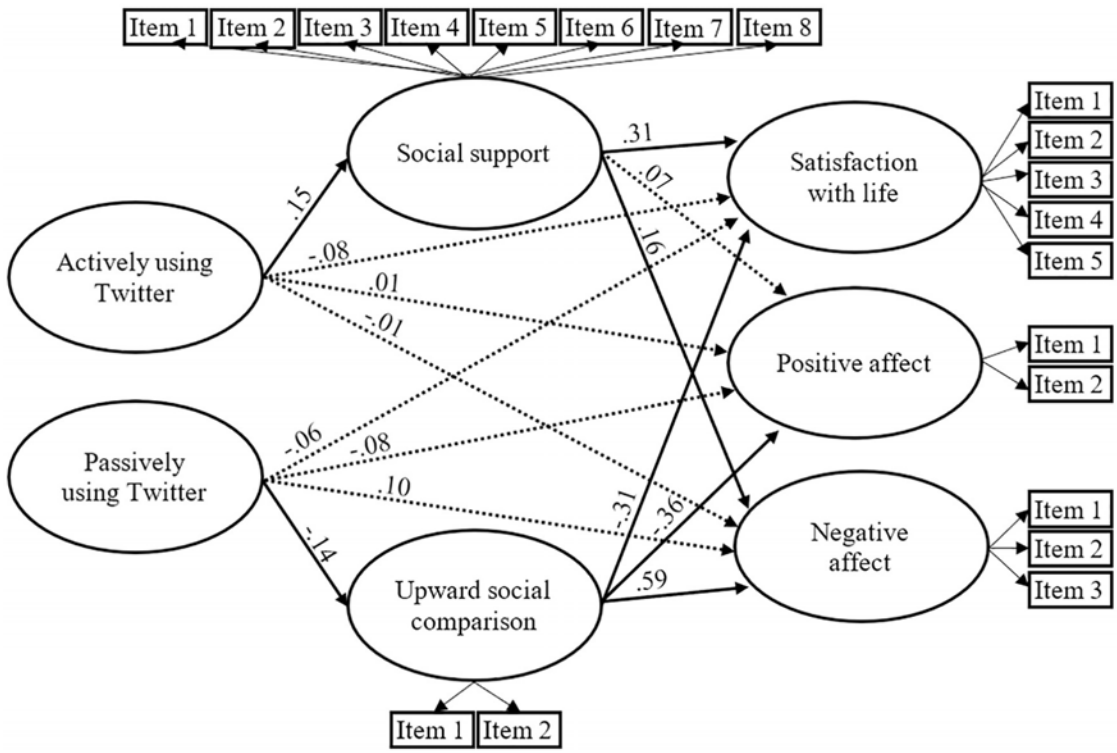


Fig 3. The estimated standardized parameters of the Twitter model. Dashed lined indicate nonsignificant paths ($p > .05$). The three components of well-being were controlled–but not displayed—for gender and age: Age was associated with satisfaction with life (β = .26, $p < .05$), negative affect (β = -.33, $p < .05$), and positive affect (β = .26, $p < .05$); women had less positive affect (β = -.30, $p < .05$), and more negative affect (β = .14, $p < .05$).

https://doi.org/10.1371/journal.pone.0248384.g003

through upward social comparison [20]. This study extends the scope of this conclusion by systematically testing the model to various SNSs (Facebook, Instagram, Twitter, TikTok) within a wider context: the COVID-19 pandemic.

First of all, participants' increase in the use of all SNSs during the quarantine measures strengthens the need to explore the relation between SNSs and well-being. Consistent with Verduyn et al. [20], upward social comparison mediated the negative association between passively using Facebook and well-being. Nevertheless, no relation was found for active Facebook usage and social support (hypothesis 1 partially supported). Instagram showed the opposite relation: social support mediated the positive association between actively using Instagram and well-being (satisfaction with life and negative affect). In contrast to Chae [22], no relation was found for passively using Instagram and upward social comparison (hypothesis 2 partially supported). One surprising outcome is that negative affect were positively related to social support and using actively Facebook. However, in line with Rimé, Bouchat, Paquot and Giglio [40], it is plausible that interacting with others on SNSs elicits emotional reactivation rather than discharge. As a consequence, obtaining social support during the COVID-19 pandemic, a negative and painful event, may increase negative affect. This result is particularly interesting and highlights the role of the socio-emotional context in the relation between SNSs and well-being. As regard to TikTok, no association with well-being, social support or upward social comparison was found (hypothesis 3 not supported). Finally, actively using Twitter was associated with more social support, and passively using Twitter with less upward social comparison.

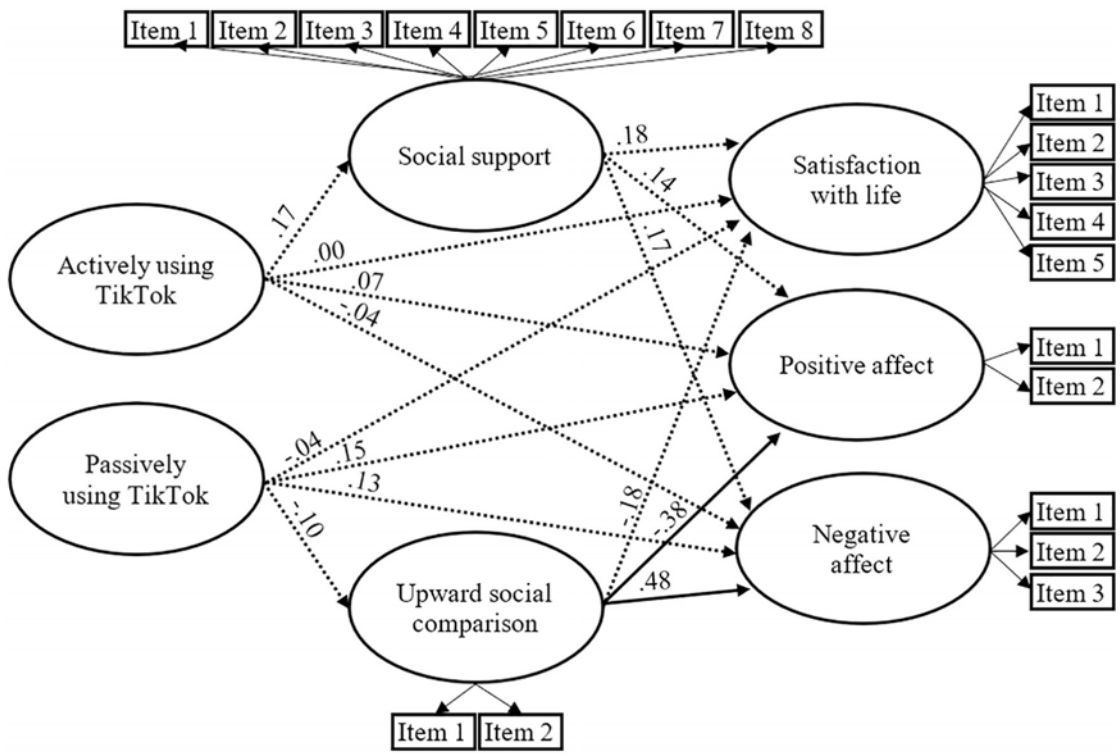


Fig 4. The estimated standardized parameters of the TikTok model. Dashed lined indicate nonsignificant paths ($p > .05$). The three components of well-being were controlled–but not displayed—for gender and age: women had less positive affect ($\beta = -.30$, $p < .05$), and more negative affect ($\beta = .26$, $p < .05$).

https://doi.org/10.1371/journal.pone.0248384.g004

Furthermore, social support mediated the relation between using actively Twitter and satisfaction with life, and upward social comparison mediated the relation between passively using Twitter and negative affect (hypothesis 4 not supported). In other words, our results are fully consistent with those of Chae [22] and demonstrate that, rather than an absence of relation [21], both active and passive usage of Twitter can be positively related to well-being. Which might seem surprising—the negative association between passive usage of Twitter and upward social comparison–may find an explanation in the social context of Twitter. Indeed, previous studies have shown that negative messages are shared faster on Twitter [41] and that popular events on Twitter are associated with negative emotions [42]. Recently, Waterloo, Baumgartner, Peter and Valkenburg [43] showed than negative emotions are perceived as more appropriate on Facebook and Twitter, compared to Instagram. Hence, it is plausible that Twitter's users scrolling through their Twitter news feed and seeing constant bad news from their followers, are more inclined to compare their situation with what they consider to be worse (i.e. downward social comparison), rather than better off (i.e. upward social comparison). Conversely, Facebook is known to be a place for positive self-presentation and impression management [26], which could explain the positive association with upward social comparison.

In that respect, it seems that the model proposed by Verduyn et al. [20] does not stand for every kind of SNSs. Facebook and Instagram use matched partially to the underlying mechanisms, but TikTok use had almost no relation to well-being, and passive usage of Twitter was negatively associated with upward social comparison. The issue is therefore to understand what characteristics and features of SNSs are accountable for these differences. In contrast with Pittman and Reich [21], the findings did not support the architecture of SNSs. Rather, it

seems that users' motivations are more indicative: Twitter and TikTok use during quarantine were not related with social relationships, contrary to Facebook and Instagram. But while TikTok use was only related to entertainment, Twitter use was also related to some purposive values which are considered as a subtype of social support [i.e., informational support, 32]. This is, by the way, fully in line with the literature on motivations to use Facebook, Instagram, Twitter, and TikTok [26–30]. Future studies should further explore how users' motivations affect the relation between SNSs and well-being.

Concerning the specific context of the COVID-19 pandemic, in the words of IJzerman et al. [44] "psychological science is not yet a crisis-ready discipline", and caution should therefore be taken to give recommendations. This single study does not allow to give advice. Maybe conclusions are solely to not systematically promote an overall use of SNSs but rather to distinguish active and passive usages, and to differentiate social network sites due to their specificities.

Finally, the present study is not devoid of limitations. Since participants were recruited via academic mailing lists from social science, the sample is quite biased towards academic people, as well as women (there are a majority of women in social science). This kind of limitation is common in research about social network sites, but we could suspect that this unbalance-sample limits the generalization of the results. Likewise, the small number of participants having a TikTok account, and the fact that all participants were Francophone, highlight the need to replicate the study in other populations. Second, to avoid demotivating respondents, we have limited the questionnaire length. Consequently, passive and active SNS usages have been measured with one item. Although they are considered as separated constructs in the literature [13], we may suspect that the use of single items has increased their association. Future studies should therefore assess specific activities on each SNS. For the same reason, only three kinds of motivation were included. But there is a lot of other reasons to use social network sites, like self-enhancement or self-documentation. Thirdly, as noted by an anonymous reviewer, we think that another good way of assessing our hypotheses would have been to test a model including all social network simultaneously. However, we think that this kind of modeling requires much more participants to draw valid conclusions. Last but not least, this study is cross-sectional, which do not allow to speak in terms of causality or consequences. For example, this study cannot support if people with lower well-being go on SNSs to increase their social support [45]. Future studies should therefore employ longitudinal and experimental designs.

Conclusions

The current research addresses the complex relation between SNSs and well-being. It extends the literature on passive and active usages by opening the reflection on various kinds of SNS. Passive usage of Facebook was related to social comparison, which, in turn, was associated with lower well-being. Besides, active usage of Instagram was related to social support, which, in turn, was associated with greater satisfaction with life but also negative affect. Regarding Twitter, active usage was also related to social support, which, in turn, was associated with greater satisfaction with life; but passive usage was rather negatively associated with upward social comparison, which, in turn, was associated with more negative affect. In contrast, TikTok use was not associated with well-being. Taken together, this study demonstrates that the differences between SNSs must be considered to truly investigate how SNSs shape human interactions—generalization to every kind of SNS should always be undertaken with caution.

Acknowledgments

We thank the participants for having dedicated their precious time to our study, especially in these difficult times.

Author Contributions

Conceptualization: Alexandra Masciantonio, David Bourguignon, Pierre Bouchat, Manon Balty, Bernard Rimé.

Data curation: Alexandra Masciantonio, Pierre Bouchat.

Formal analysis: Alexandra Masciantonio, Pierre Bouchat.

Investigation: Alexandra Masciantonio, David Bourguignon, Pierre Bouchat, Manon Balty, Bernard Rimé.

Methodology: Alexandra Masciantonio, David Bourguignon, Pierre Bouchat, Manon Balty, Bernard Rimé.

Project administration: David Bourguignon, Bernard Rimé.

Resources: Alexandra Masciantonio, David Bourguignon, Pierre Bouchat, Bernard Rimé.

Software: Alexandra Masciantonio, Pierre Bouchat.

Validation: Alexandra Masciantonio.

Visualization: Alexandra Masciantonio.

Writing – original draft: Alexandra Masciantonio.

Writing – review & editing: Alexandra Masciantonio, David Bourguignon, Pierre Bouchat, Bernard Rimé.

References

1. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Social Science & Medicine. 2000 Sep 15; 51(6):843–57. https://doi.org/10.1016/s0277-9536(00)00065-4 PMID: 10972429
2. Jetten J, Haslam SA, Haslam C. The case for a social identity analysis of health and well-being. In: Jetten J, Haslam SA, Haslam C, editors. The social cure: Identity, health and well-being. Psychology Press; 2012. p. 1–19.
3. Altena E, Baglioni C, Espie CA, Ellis J, Gavriloff D, Holzinger B, et al. Dealing with sleep problems during home confinement due to the COVID-19 outbreak: Practical recommendations from a task force of the European CBT-I Academy. Journal of Sleep Research. 2020; 29(4):e13052. https://doi.org/10.1111/jsr.13052 PMID: 32246787
4. Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y. A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. Gen Psychiatr [Internet]. 2020 Mar 6 [cited 2020 Sep 2]; 33(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7061893/. https://doi.org/10.1136/gpsych-2020-100213 PMID: 32215365
5. Casey N. How social networks can do good while we're all trapped indoors. The Verge [Internet]. 2020 Mar 13; https://www.theverge.com/interface/2020/3/13/21176880/covid-19-quarantine-social-distancing-isolation-loneliness-zoom-fortnite.
6. Chamberlain J. Coronavirus has revealed the power of social networks in a crisis. The Conversation [Internet]. 2020 May 12; https://theconversation.com/coronavirus-has-revealed-the-power-of-social-networks-in-a-crisis-136431.
7. Ellison NB, boyd D. Sociality through social network sites. In: Dutton WH, editor. The Oxford handbook of Internet studies. Oxford, UK: Oxford University Press; 2013. p. 151–172.
8. Diener E, Heintzelman SJ, Kushlev K, et al. Findings all psychologists should know from the new science on subjective well-being. Canadian Psychology/Psychologie canadienne. 2017; 58(2):87–104.
9. Lin LY, Sidani JE, Shensa A, Radovic A, Miller E, Colditz JB, et al. Association between social media use and depression among US young adults. Depression and anxiety. 2016; 33(4):323–331. https://doi.org/10.1002/da.22466 PMID: 26783723
10. Sagioglou C, Greitemeyer T. Facebook's emotional consequences: Why Facebook causes a decrease in mood and why people still use it. Computers in Human Behavior. 2014 Jun 1; 35:359–63.

11. Apaolaza V, Hartmann P, Medina E, Barrutia JM, Echebarria C. The relationship between socializing on the Spanish online networking site Tuenti and teenagers' subjective wellbeing: The roles of self-esteem and loneliness. Computers in Human Behavior. 2013 Jul 1; 29(4):1282–9.

12. Valenzuela S, Park N, Kee KF. Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation. J Comput Mediat Commun. 2009 Jul 1; 14 (4):875–901.

13. Gerson J, Plagnol AC, Corr PJ. Passive and Active Facebook Use Measure (PAUM): Validation and relationship to the Reinforcement Sensitivity Theory. Personality and Individual Differences. 2017 Oct 15; 117:81–90.

14. Burke M, Marlow C, Lento T. Social network activity and social well-being. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2010. p. 1909–1912.

15. Kim J, Lee J-ER. The Facebook Paths to Happiness: Effects of the Number of Facebook Friends and Self-Presentation on Subjective Well-Being. Cyberpsychology, Behavior, and Social Networking. 2011 Nov 30; 14(6):359–64. https://doi.org/10.1089/cyber.2010.0374 PMID: 21117983

16. Shaw AM, Timpano KR, Tran TB, Joormann J. Correlates of Facebook usage patterns: The relationship between passive Facebook use, social anxiety symptoms, and brooding. Computers in Human Behavior. 2015 Jul 1; 48:575–80.

17. Shi Y, Yue X, He J. Understanding social network sites (SNSs) preferences: personality, motivation, and happiness matters. In: International Conference on Online Communities and Social Computing. Springer; 2013. p. 94–103.

18. Tandoc EC, Ferrucci P, Duffy M. Facebook use, envy, and depression among college students: Is facebooking depressing? Computers in Human Behavior. 2015 Feb 1; 43:139–46.

19. Verduyn P, Lee DS, Park J, Shablack H, Orvell A, Bayer J, et al. Passive Facebook usage undermines affective well-being: Experimental and longitudinal evidence. Journal of Experimental Psychology: General. 2015; 144(2):480–8. https://doi.org/10.1037/xge0000057 PMID: 25706656

20. Verduyn P, Ybarra O, Résibois M, Jonides J, Kross E. Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review. Social Issues and Policy Review. 2017; 11(1):274–302.

21. Pittman M, Reich B. Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. Computers in Human Behavior. 2016; 62:155–167.

22. Chae J. Reexamining the relationship between social media and happiness: The effects of various social media platforms on reconceptualized happiness. Telematics and Informatics. 2018 Sep 1; 35 (6):1656–64.

23. Clement J. Global social networks ranked by number of users 2020 [Internet]. Statista. 2020 [cited 2020 May 4]. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

24. Mohsin M. 10 TikTok Statistics That You Need to Know in 2020 [Infographic] [Internet]. Oberlo. 2020 [cited 2020 May 4]. https://wp-en.oberlo.com/blog/tiktok-statistics.

25. Bossetta M. The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 US election. Journalism & mass communication quarterly. 2018; 95(2):471–496.

26. Nadkarni A, Hofmann SG. Why do people use Facebook? Personality and individual differences. 2012; 52(3):243–249. https://doi.org/10.1016/j.paid.2011.11.007 PMID: 22544987

27. Sheldon P, Bryant K. Instagram: Motives for its use and relationship to narcissism and contextual age. Computers in human Behavior. 2016; 58:89–97.

28. Johnson PR, Yang S. Uses and gratifications of Twitter: An examination of user motives and satisfaction of Twitter use. In: Communication Technology Division of the annual convention of the Association for Education in Journalism and Mass Communication in Boston, MA. 2009.

29. Park CS. Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement. Computers in Human Behavior. 2013; 29(4):1641–1648.

30. Omar B, Dequan W. Watch, Share or Create: The Influence of Personality Traits and User Motivation on TikTok Mobile Video Usage. International Journal of Interactive Mobile Technologies (iJIM). 2020 Mar 12; 14(04):121–37.

31. Cheung CM, Chiu P-Y, Lee MK. Online social networks: Why do students use facebook? Computers in Human Behavior. 2011; 27(4):1337–1343.

32. Nick EA, Cole DA, Cho S-J, Smith DK, Carter TG, Zelkowitz RL. The Online Social Support Scale: Measure development and validation. Psychological assessment. 2018; 30(9):1127–43. https://doi.org/10.1037/pas0000558 PMID: 29781664

33. Brunot S, Juhel J. Comparaisons sociales et temporelles, estime de soi et activité de recherche d'emploi en situation de chômage de longue durée. LAnnee psychologique. 2012; 112(2):197–226.

34. Fredrickson B. Positivity: Top-Notch Research Reveals the 3-to-1 Ratio That Will Change Your Life. Harmony; 2009. 288 p.

35. Gaudreau P, Sanchez X, Blondin J-P. Positive and Negative Affective States in a Performance-Related Setting. European Journal of Psychological Assessment. 2006 Jan 1; 22(4):240–9.

36. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. Journal of Personality Assessment. 1985 Feb 1; 49(1):71–5. https://doi.org/10.1207/s15327752jpa4901_13 PMID: 16367493

37. JASP Team. JASP (Version 0.12.2) [Computer software]. 2020.

38. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software. 2012 May 24; 48(1):1–36.

39. Gana K, Broc G. Introduction à la modélisation par équations structurales: Manuel pratique avec lavaan. ISTE Group; 2018.

40. Rimé B, Bouchat P, Paquot L, Giglio L. Intrapersonal, interpersonal, and social outcomes of the social sharing of emotion. Current Opinion in Psychology. 2020 Feb 1; 31:127–34. https://doi.org/10.1016/j.copsyc.2019.08.024 PMID: 31590107

41. Naveed N, Gottron T, Kunegis J, Alhadi AC. Bad news travel fast: A content-based analysis of interestingness on twitter. In: Proceedings of the 3rd international web science conference. 2011. p. 1–7.

42. Thelwall M, Buckley K, Paltoglou G. Sentiment in Twitter events. Journal of the American Society for Information Science and Technology. 2011; 62(2):406–418.

43. Waterloo SF, Baumgartner SE, Peter J, Valkenburg PM. Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. new media & society. 2018; 20(5):1813–1831.

44. IJzerman H, Lewis NA, Weinstein N, DeBruine LM, Ritchie SJ, Vazire S, et al. Psychological Science is Not Yet a Crisis-Ready Discipline [Internet]. PsyArXiv; 2020 Apr [cited 2020 May 6]. https://osf.io/whds4.

45. Trepte S, Scharkow M. FRIENDS AND LIFESAVERS How Social Capital and Social Support Received in Media Environments Contribute to Well-Being. In: Reinecke L, Oliver MB, editors. The Routledge Handbook of Media Use and Well-Being: International Perspectives on Theory and Research on Positive Media Effects. New-York: Routledge; 2016. p. 304–16.

PX0323

Check for updates

Article

SM+S
social media + society

Social Media + Society
January-March 2022: 1–11
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20563051221086241
journals.sagepub.com/home/sms
SAGE

Why's Everyone on TikTok Now? The Algorithmized Self and the Future of Self-Making on Social Media

Aparajita Bhandari[1*] and Sara Bimo[2*]

Abstract

The video-sharing social media platform TikTok has experienced a rapid rise in use since its release in 2016. While its popularity is undeniable, at the first glance, it seems to offer features already available on previously existing and well-established platforms such as Instagram, YouTube, and Facebook. To understand processes of self-making on TikTok, we undertake two methods of data collection: a walkthrough of the app and its surrounding environment, and 14 semistructured participant interviews. A qualitative analysis of this data finds three distinct themes emerge: (1) awareness of the algorithm, (2) content without context, and (3) self-creation across platforms. These results show that TikTok departs from existing platforms in the model of self-making it engenders, which we term "the algorithmized self"—a complication of the pre-existing "networked self" framework.

Keywords

social media, TikTok, algorithms, self-making, self-representation

Introduction

Since its release in 2016, the video-sharing platform TikTok has enjoyed a meteoric rise in popularity: as of February 2021, it has been downloaded over 2.6 billion times worldwide (with 315 million of these downloads occurring in the first quarter of 2020), and has approximately one billion monthly active users (TikTok Statistics, 2021). Originally released by the Chinese company ByteDance under the name "Douyin" (as it is still known in China), TikTok only became available worldwide after merging with another Chinese social media service, Musical.ly, in 2018. Centered around video sharing, TikTok allows users to create music and lip-sync videos of 2–15 s, and looping videos of initially 2–60 s (Williams, 2020) with recent changes now allowing verified users to create videos that are up to 3 min long (TikTok Newsroom, 2021). It also allows users to "duet" (a function where users can directly respond to an existing video and display the response and original side by side), "stitch" (clip and integrate other users' video into one's own), and live-stream content, in addition to many other features. TikTok is a constantly evolving and dynamic platform, with new features regularly being added and changed, often rapidly. While it is perhaps best known for its music- and dance-oriented videos, the platform features a diverse variety of content, from humorous skits to beauty tutorials to political advertisements to educational DIY videos. As the New York Times states, "[On TikTok] You can join a dare-like challenge, or participate in a dance meme, or make a joke. Or you can make fun of all of these things. TikTok assertively answers anyone's what should I watch with a flood" (Herrman, 2019).

While its rise in popularity is noteworthy, it may be slightly surprising to those only peripherally aware of the platform. At first glance, the particular services offered by TikTok seem to be similar to previously established social media platforms. Functionally, it can be understood as the latest iteration of a growing trend in social media which has seen the popularization of video creation and sharing in the form of live-streaming and short looping videos. It is regarded by many as the spiritual successor of the now-defunct app Vine, which was

[1]Cornell University, USA
[2]York University, Canada

*Co-first authorship

Corresponding Author:
Aparajita Bhandari, Cornell University, 100 Mann Drive, Ithaca, NY 14853-0001, USA.
Email: ab2725@cornell.edu

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

based on sharing videos of less than 6 s and which fostered a distinct brand of surrealist "millennial" humor that is echoed in the creative chaos that characterizes many TikTok videos (Anderson, 2020). TikTok also shares similarities with other popular social media: both Instagram and Facebook allow live streaming, and since 2016 have added a "stories" feature which allows users to create short looping videos of less than 20 s (a functionality which had been available on Snapchat since 2013) (Verstraete, 2016). In addition, the video-sharing giant YouTube, released in 2005, has long offered the ability to share short videos and added live-streaming in 2008 (Anderson, 2020). Such platforms, which are predicated primarily on the sharing of visual media, have been studied extensively in terms of the sociality and self-making processes they engender for their users (Thumim, 2012; Tiidenberg & Whelan, 2017).

However, there is one key new element that sets TikTok apart from other outwardly similar social media platforms: the prevalence of "the algorithm." TikTok unprecedentedly centers algorithmically driven feeds and algorithmically driven experiences. On TikTok, unlike on other platforms, the user experience is obviously, unambiguously, and explicitly driven by what is commonly called the "For You" algorithm (Xu et al., 2019). While algorithms are becoming increasingly prevalent across the social media landscape, on other platforms, they are still ostensibly only an "element" (or enhancement) of an otherwise user-driven experience. Of the major social media platforms on the market, TikTok is the only one to position its algorithm at the center of the social experience it engenders; the algorithm determines the type of video content the user is exposed to, and viewing this content makes up the majority of the experience on the platform.

This key element—the forefronted algorithm—of TikTok has the potential to alter previously seen models of self-making on social media. While there has been extensive research that has examined self-making on social media, especially through the creation and consumption of visual content, and work that has probed into the connection between algorithms and user self-creation, these two areas of research have not often been put into conversation. This article seeks to fill this gap in the literature by pursuing the following research question: how is "the self" created on and through TikTok, and does the process of self-making on TikTok differ when compared to other social media sites?

To address the particular issue of "self-making" on TikTok and to accurately draw comparisons to other social media platforms, it is first necessary to review the extant literature regarding issues related to "the self" on social media platforms and establish an understanding of the common model of self-making on social media. In the next section, we explore what the current literature says about how self-making "works" on social media, and how algorithmic experiences intersect with self-making practices.

Literature Review

The Self and Self-Representation

When it comes to examining the creation of the self on social media, the concept of "self-representation" is highly relevant. Self-representation can be understood as a form of mediation and involves the social process of creating media objects that stand in for, transform, and recreate a person or object (Thumim, 2012).

While people have used technology and media to record, share, and communicate the self since time immemorial, the tools and environments furnished by social media spaces have supplemented, and arguably, propagated self-representational practices (Rettberg, 2014). Visual media sharing has become a "hallmark of contemporary internet culture" (Mirzoeff, 2016). For example, Facebook users, upload 350 million photos per day; users of Snapchat and Whatsapp share over 700 million daily photo shares (Tiidenberg & Whelan, 2017). On social media, visual self-representations "circulate between individuals and groups, facilitate the creation and maintenance of relationships, memories, norms, and ideologies, and are widely understood as tools for identity formation and communication" (Tiidenberg & Whelan, 2017, p. 141; Van Dijck, 2008).

How Does Self-Representation Work?

When it comes to the extant literature on self-representation, traditional cultural studies frameworks offer a model of how people self-represent with visual media: visual representations can be understood as indexical images that stand in for or represent a person or object (Hall, 1997, p. 1). However, conceiving of images as merely indexical representations of what they depict disregards the social contexts within which online images are embedded (Cruz & Thornham, 2015; Frosh, 2015), and risks overemphasizing the media itself as absolutely central to the meaning production process and thus understating the technocultural processes which intervene in the creation of media texts and the circulation of meaning in society (Thumim, 2012, p. 54). Thus, scholars have called for more nuanced understandings of self-representation that move away from indexicality as the primary model for how we produce and circulate meaning.

This issue has been addressed by authors such as Thumim, who proposes that self-representation be understood as a genre rather than a discrete set of practices (Thumim, 2012). Genre in this case refers to something multidimensional and intertextual, something which can be understood as a "tacit agreement" between producers and audiences (p. 166). Content within the genre of self-representation will contain combinations of certain elements, such as community, experience, interior worlds, emotion, personal artifacts, and so on, but no one work is necessarily bound to contain all the

elements associated with this genre. Furthermore, examples in the genre may differ strongly in their "politics, purpose, conditions of production, and more" (p. 167). Shifting our understandings away from indexicality in this manner brings into relief the social, context-dependent, and mediated qualities of self-representation.

Tiidenberg and Whelan (2017) build on the work of Thumim to further complicate our understanding of self-representation. Departing from media and cultural scholarship, they point out that self-representation can also be understood as a process of "identification" (in the psychoanalytical sense). Psychoanalytical identification is defined as the psychological process of association between oneself and something else (Tiidenberg & Whelan, 2017). Thus, self-representation-as-identification involves the construction of the self in terms of the objects and images that we identify with (Bamberg, 2011).

A psychoanalytical lens re-centers "the self" in self-representation. Emphasizing the self and reorienting conceptualizations away from static representations to dynamic processes of self-making and identification also gives rise to a more nuanced interpretation of identity. It allows identity to be something that is often multiple, fragmented, and discordant; constructed across varying discourses, practices, and positions, and constantly in the process of change and transformation (Hall, 1997).

Within this article, we draw on the notions of genre and self-representation as proposed by Thumim (2012) and Tiidenberg and Whelan (2017), respectively, as an entry point into understanding how self-making occurs on social media. We use the term self-making to highlight the set of intertextual and flexible practices, conventions, and norms of both the production and consumption of visual content that constitutes identity creation online.

The Networked Self

Within the context of online self-making research, the "networked self" model examines how social media spaces constitute sites of self-presentation and identity management (Papacharissi, 2011). Papacharissi (2013) writes that the "appeal" of social media sites is their provision of a "stage for self-presentation and social connection" (p. 206): through their multimedia capabilities, they provide props (in the form of text, photo, video, etc.) that facilitate self-presentation. In such an environment, the self is performed through public displays of social connections: thus, the self is created through the "reflexive process of fluid association with social circles" (Papacharissi, 2013, p. 208). In this school of thought, social media sites are ultimately projects in managing sociality and negotiating self-expression through these social ties: one self-represents by engaging with one's network within the context of the social media sites.

This understanding of a "networked self" posits that social media sites—apart from the influence they may have on social practice through their design and affordances—are relatively neutral stages, or tools, for individuals to engage in these processes of identity management and representation. Interaction and curation are ostensibly self-directed; despite the growing prevalence of algorithmically directed feeds, for example, these sites at least claim that the content that users consume is essentially within their control and is ultimately chosen by the user through their conscious choices: the site thus provides a space for free self-representation (Van Dijck, 2013). The conceit of these social media platforms is the notion that technology is mainly a tool or extension of the user's will: digital environments merely supplement existing modes of social experience, as opposed to fundamentally shaping or altering it (Papacharissi, 2013).

In the context of visually based social media platforms, the notion of the networked self operates hand in hand with the previously discussed concept of self-making: users construct their identities by engaging with (through varied and multidimensional practices such as identification, indexicality, etc.) the visual media objects created and posted by others in their network.

Algorithms and Identity

The networked self-model operates on the assumption that social media spaces are merely neutral stages for this "reflexive process of fluid associations with our social circles." However, in recent years, this notion has been increasingly troubled due to the growing prevalence of algorithmically directed feeds and social media experiences. Scholars have begun to question the impact that such new forms of technology can have on self-making processes.

Hearn (2010) argues that under current conditions of modernity, self-hood, and self-presentation are inextricably entangled with advances of capitalism and its ever-evolving search for new forms of profit. Therefore, it follows that the self must be understood in relation to the "datalogical turn," wherein the capitalist mode of production that characterizes our society has become predicated on the collection, generation, and sale of vast amounts of consumer data (Gregory et al., 2015). This new form of "effective" capitalism takes as its raw material the "desires, emotions, and forms of expressivity" that users present in digital spaces; thus, the practices involved in online self-making become part of a wider economic infrastructure (Hearn, 2017, p. 63).

Affective capitalism is enabled by tools such as algorithms and automated recommendation systems. These technologies allow for the self-making practices of users to be collected and sold by companies like Twitter and TikTok, and, through processes of aggregation, abstraction, and categorization, be rendered into consumer profiles. The categories that are derived from user's online actions are ultimately

projected back onto them, enframing them in an "algorithmic identity—an identity formation that works through mathematical algorithms to infer categories of identity on otherwise anonymous beings" (Cheney-Lippold, 2011, p. 165 as seen in Hearn, 2017).

Through algorithms, affective capitalism succeeds in enacting particularly insidious and far-reaching forms of control: control over user identity. Cheney-Lippold sheds light on the implications such technologies have over self-making practices, arguing that "the automated categorization practices and the advertisements and content targeted to those categorizations effectively situate and define how we create and manage our own identities" (Cheney-Lippold, 2011, p. 177).

In summation, the current literature on the impact of algorithms on self-making argues that algorithms have the power to define and situate our identities by fitting users into predefined categorization schema for the purpose of data gathering and advertising. However, this literature tends to underscore or even overlook the agency of users within these processes of self-making. The current study seeks to bridge the gap in the literature between studies of self-representation on platforms that can be said to follow the "networked self" model of self-making, which ascribes a relatively high degree of agency to individual users, and the literature that interrogates the effects that increased algorithmization may have on user identity creation on social media (which highlights the reduced agency users experience in terms of self-making practices). The central questions guiding this project are: how do algorithms influence the process of "self-making" on social media? How do everyday users make sense of their experiences on algorithmic social media streams?

We answer these questions by exploring what new models of self-making might result from the combination of an otherwise traditional social media platform with the heightened presence of the algorithm. We take up the issues of self-hood and self-making under the influence of the algorithm by taking a look at user practices on TikTok, a platform where user experience is heavily guided by an algorithm.

Methods

To conduct a rigorous inquiry, we employed two means of data collection in this project. First, we used the walkthrough method as described by Light et al. (2018). This method, grounded in Actor Network Theory, involves the systematic collection of data throughout various steps of app registration and entry, everyday use, and discontinuation of use to analyze an app in its entirety.

In accordance with this method, our walkthrough data collection process comprised two distinct phases. First, the environment of expected use was established by examining the app store descriptions, white papers, and newsroom documents put forward by the app from its launch until February 2020. This was taken in conjunction with narratives put forward in media articles about TikTok. In the second phase, we conducted a technical walkthrough of TikTok. This was done in February 2020 on two different phones, using both Android and iPhone versions of the app. To analyze the everyday use of the app both researchers used TikTok for at least 30 min daily for a period of 1 month in 2020, and collected extensive fieldnotes, screenshots, and screen recordings. This allowed us to investigate the environment of expected use, user interface arrangement, functions and features, textual content and tone, and symbolic representations present in the app (Light et al., 2018).

In addition to collecting data through the walkthrough method, we conducted in-depth interviews with 14 college students who use TikTok, ranging in age from 18 to 24 years. College students are not necessarily a representative sample of all those who use TikTok, and the findings of this article cannot be generalized beyond the scope of this study. However, it is important to note that almost 60% of TikTok users are under 30 with roughly 41% of users between the ages of 16 and 24 years (TikTok Statistics, 2021).

Our interview participants were from a US east coast university and came from varied educational backgrounds: their completion levels (freshman through postgraduate) and courses of study included computer science, communication, engineering, and hotel management, among other subjects. These participants were recruited and compensated through the university's research management system. The interviews were all conducted over video conferencing technology, and upon completion, participants were compensated with a university research credit. Women were greatly overrepresented in our sample ($n = 11$), which may reflect larger social media usage patterns (Anderson & Jiang, 2018). All of our interviewees were enrolled in university, reflecting a certain level of relative privilege.

The interviews had an average duration of 35–45 min, and followed a semistructured protocol. Discussion topics included initial use and changing experiences with TikTok, regular usage patterns, the types of content viewed and shared and why, and participants' likes and dislikes of the platform, and forms of self-expression (e.g., "Do you often post TikToks that you think represent you?"). Interviews were recorded with permission and transcribed, and the researchers used an inductive approach to establish coding categories. We followed a grounded theory approach, which involves subjecting "inductive data to rigorous comparative analysis that successively moves from studying concrete realities to rendering a conceptual understanding from these data" (Charmaz & Belgrave, 2012, p. 347). The process of data collection and analysis was overall iterative and inductive; for example, when we heard about the different "sides"' of Tik Tok, we added more follow-up questions about this to our interview protocol.

Findings and Analysis

Upon analysis of our data, we found that three overarching themes emerged. We present the findings of our walkthrough and interview data in tandem through a discussion of these themes, which are as follows: (1) awareness of the algorithm, (2) content without context, and (3) self-creation across platforms. Taken together these themes inform what we term the "algorithmized self" on TikTok, a conceptualization of the changing nature of self-making online.

Theme 1: Awareness of the Algorithm

The predominant theme that emerged from both the participant interviews and the walkthrough analysis was a sense of heightened awareness of the "infrastructure" of TikTok: in particular, the algorithm that curates a user's home feed. Participants often referred to this as the "For You" algorithm as it determines the content that appears on the For You page, which consists of an endless stream of videos uniquely personalized for each user. While the exact workings of the algorithm are kept secret by ByteDance, the basic mechanics are such: TikTok considers data from user interactions (such as videos liked and shared, accounts followed, comments, etc.), video information (captions, sounds, and hashtags), and device and account settings (language preference, country setting, and device type) to present users with the videos that are most relevant to their interests (Latermedia, 2020).

All participants interviewed discussed the For You algorithm at length, and indicated that their engagement with it constituted a significant portion of their overall activity on the platform. In many cases, participants stated that the accuracy of the algorithm was a significant draw, and was the reason for either their initial interest in or continued use of TikTok: "It is such a good algorithm, I have no idea how they do it. Probably a lot of data collection, I'm assuming."

Findings from the walkthrough analysis suggested that this hyper-awareness of the algorithm is augmented by the design of TikTok. Upon opening the application, users are directed toward the For You page, which constitutes the "home page" of the platform and from where all other activity types can be found. While many social media platforms have home pages that display a "feed" of content that comprises the posts made by followed users, pages, or friends, they additionally draw attention to other "spaces" of the site: for example, while both Twitter and Facebook situate users in their home feed, to begin with, they also direct users toward other activities, such as looking at profiles, sending messages, and so on through a design which places visual prominence on a variety of interaction types. By contrast, the visual design of TikTok obfuscates the interaction types which make up the bulk of social activity in digital spaces (i.e., activities that connect users). The videos that make up the For You page feed take up the entire screen, and it can be difficult to navigate out of this space to find the relatively small icons that will allow users to comment on videos, follow others, send messages, and so on. Visually, such activities are presented as secondary to the content presented by the algorithm.

This design decenters the traditionally "social" activity on the platform: for example, users need not be following or otherwise engaging with other creators to see their videos. Rather, they need only interact with the For You algorithm for a certain period of time in order for it to "get to know" their personalities and interests well enough to present them with accurate and entertaining content: the algorithm does the work that "following" does on other platforms, but more effectively and efficiently.

The For You page is visually and organizationally dominant in the TikTok experience, and this dominance reflects its importance and prominence in the minds of users, who spend the bulk of their time on the platform experiencing and engaging with this page. The prominence of the For You page leads naturally to the prominence of the For You algorithm in the psyche of users, as they are quickly made aware that their experience on TikTok is almost entirely shaped by this algorithm. In centering the algorithm in this way, TikTok differs from other mainstream social media sites, which frequently employ algorithms to curate user content, but often do so to supplement other user interactions (e.g., Facebook uses algorithms to help curate user feeds [and promote targeted advertisements], but maintains the conceit that users design their own feeds through their choices to follow certain friends or pages) (Willson, 2017). While other platforms obfuscate the fact of algorithmic intervention, TikTok dispenses with this illusion by highlighting the role its algorithms play in shaping user experience and promoting user-generated content.

In addition to exhibiting a heightened awareness of this algorithm, many participants indicated a temporal element to their relationship with it; their experience on TikTok changed the longer they were on the platform. This often resulted in the use of distinctly personifying and humanizing language to describe the algorithm. Interviewees repeatedly described a shift in their recommended content and this shift was described as the moment that TikTok "got them" or understood them and was able to capture their personalities and interests with pinpoint accuracy. As one participant explained, "The more time I spend on TikTok, the better it gets to know my personal likes or dislikes, and it gives me more and more content that I like." Another participant described how TikTok was "boring at first," but once the algorithm "got me," the app became more entertaining. Participants expressed awareness that TikTok was collecting and sharing their personal data, but this was seen as an acceptable tradeoff for the level and quality of personalization of their feeds and recommendations provided in return. However, not all interviewees viewed this hyper-personalized curation of

content as a wholly positive feature of TikTok's algorithm; many saw the content as restricting and overly fitted to the algorithm's conceptualization of them and their tastes.

While participants discussed the algorithm at length, they were often uncertain as to exactly how it worked. At most, they understood it as an external entity that amalgamated their data input and used this information to present them with content specifically curated to their interests. One participant stated:

> When I first got on, it was a little bit more broad and exposed me to a lot of interesting and different things. Then, as the TikTok algorithm got to know me, the content is now more curated, which can be kind of annoying because if I see the same type of content over and over again, I want to sometimes get out of it, but it's hard to because I don't know how to get away from what I'm typically seeing.

Similarly, another participant stated, "The algorithm sometimes is so specific, that I just end up seeing the same video over and over again," while another expressed bafflement at the videos TikTok was presenting to them: "one week I was getting all these videos of frogs, and everyone was like, why am I on the frog side of TikTok? It was so random." Yet another participant mused about how, "there's an entire side of TikTok that I'm not exposed to at all just because of what the algorithm thinks."

The prominence of the algorithm in combination with the lack of knowledge about its inner mechanics lead to the creation of a kind of "algorithmic imaginary," an idea defined by Bucher as the "ways in which people imagine, perceive and experience algorithms and what these imaginations make possible" (Bucher, 2017). While the algorithmic imaginary is a significant part of the experience of any social media site and is likely to become even more so as algorithms gain increasing visibility and prominence in shaping online social experience, interviewees indicated that the mental energy they devoted to the TikTok algorithm far exceeded that spent on the algorithms of other social media sites.

In addition, the prevalence of the temporal element on TikTok leads to participants moving beyond simply imagining the algorithm, toward a heightened understanding of their own relationship with the algorithm; a relationship characterized as dynamic and changeable. Participants frequently took an active role in their TikTok experience, exhibiting a high degree of what we might call "algorithmic engagement." Not only were participants highly aware of the algorithm and invested in understanding it; they also interacted with it heavily (in fact, such activity often made up the bulk of their interactions on TikTok). Participants believed that they knew how exactly to interact with the affordances and activity corridors of the app to "work with" the algorithm so that it could provide them with more relevant or entertaining content.

Thus, while the prevalence of "the algorithm" as the main shaping force of user feeds does away with the illusion of user control present on other platforms (i.e., the illusion that users can completely curate the content they consume through conscious choices of whom to follow, what to like, etc.) and thus reduces the kinds of meaningful interactions that users partake in, it presents other avenues for user engagement. While other platforms ostensibly see users engaging with other people in their network through the activity corridors of the platform, TikTok sees users engaging almost exclusively with the algorithm itself. Interview participants often mentioned that they did not feel the need to follow certain creators to gain access to their content, nor did they make use of the commenting feature to discuss videos with others, as one would on other video sharing platforms. Instead, they repurposed these activity corridors as mechanisms through which they could further interact with the algorithm. As one participant explained:

> Sometimes you'll see a post where it'll just be dedicated to saying welcome to the x side of TikTok (maybe the black side or the queer side), which is obviously more of a joke. But then in the comments everyone will be talking about how to get on this side, talking about what side of TikTok we're on and how we got there. And sometimes I'll comment just to make sure that I'm engaging enough to stay on this side too if it's something I'm into.

The status of the algorithm as the main external entity with which participants interact, in combination with the personifying language participants use to describe it, results in the algorithm occupying a unique role in the TikTok experience. Rather than being merely an element or building block in the infrastructure of the platform (part of the aforementioned "stage" that a social media site provides to enable the "real" social behavior of interacting with other egos in one's network), or another independent ego with whom users can interact (as they would with a friend), the algorithm exists somewhere in between. It is an entity with which users can engage and which they can influence and manipulate, so users have some degree of control over what this algorithm shows them. However, it is largely impenetrable: it is thus not analogous to an inert "tool" with which they can freely enact their own independent processes of self-representation and identity construction. At the same time, it is not independent; it exists solely to present them with access to content that reflects their own internal worlds; content that reflects their interests, likes, and personality, and which might be seen as a curated collection representing their inner "self."

Participants occasionally displayed an awareness of the strangeness of this relationship. One participant noted with unease moments when the algorithm was almost "too accurate," describing how at one point it was presenting content that was not only reflective of their sense of humor and interests, but also their physical appearance. The participant felt

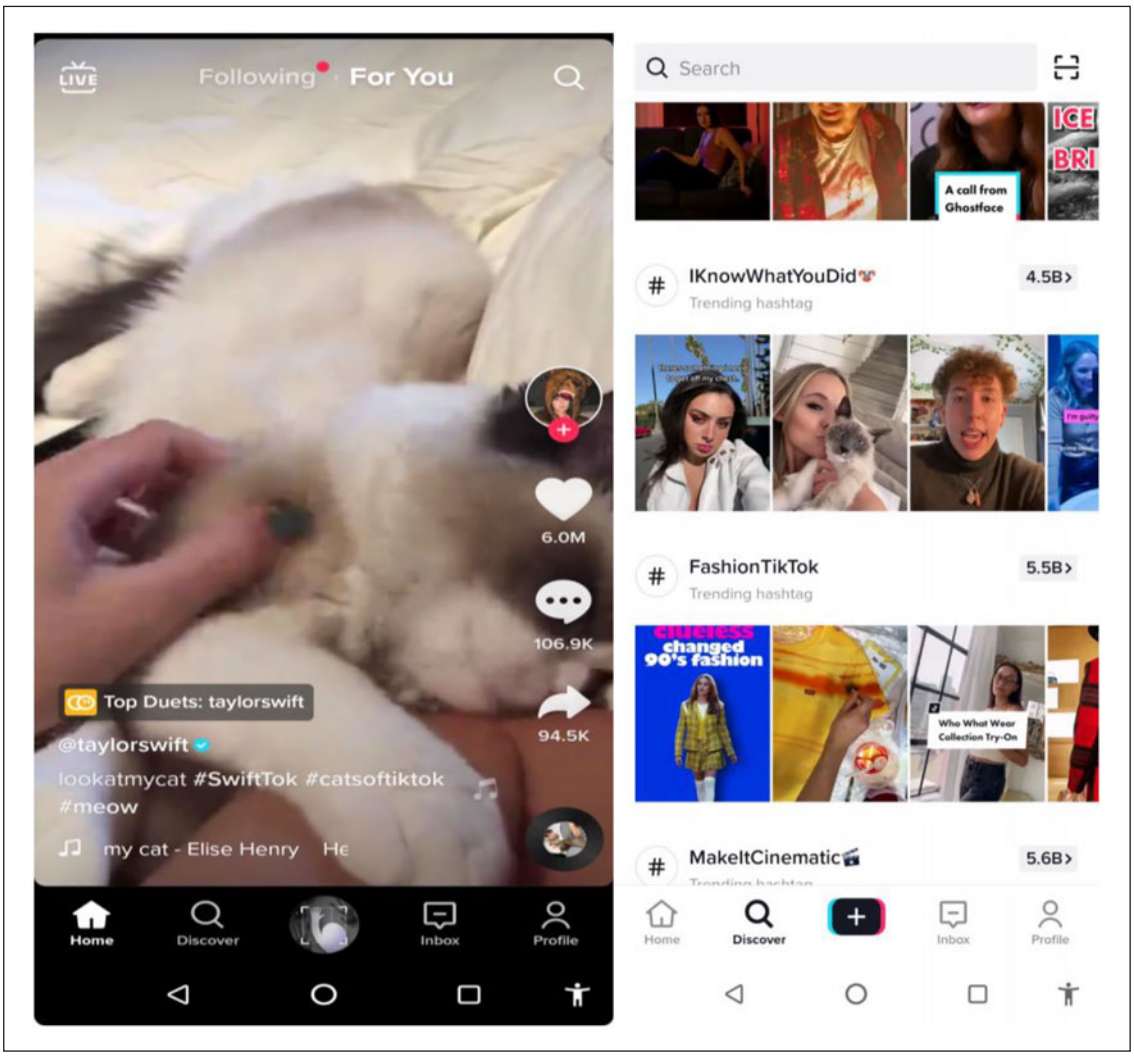


Figure 1. Screenshots of the TikTok app. On the left is an example of the For You page and on the right is the Discover tab.

"weird" about being confronted with various iterations of people who acted and looked too much like them. At this point, the participant ceased to view the algorithm's accuracy as a positive quality and felt that a line had been crossed. Other participants pointed to moments where they realized that the algorithm was showing them "repackaged" versions of themselves. These moments frequently engendered a sense of unease as participants became aware that what they had been interacting with was a version of themselves filtered through the algorithm, rather than an external ego.

Theme 2: Content Without Context

When asked to explain the specific appeal of TikTok, participants frequently praised the quality of the "content." They explained that, as a result of the accuracy of the For You algorithm (see Figure 1), they were usually highly interested in the videos curated for them and were more consistently entertained in comparison to their experiences on other social media platforms. In addition to appealing to their personal interests, participants explained that the content on TikTok was appealing in that it was a way to stay "up to date" with trends, memes, and current events. As one participant put it:

> [TikTok] just continuously recommends me content that I enjoy, like Netflix will recommend shows. I never really like any of the shows Netflix recommends me, I usually hate them [. . .] But on TikTok, the material that is recommended to me is always really good, even though it recommends such a bulk amount.

Another participant stated:

> I'm not particularly on it for [the social aspect]. I just really like the content, I feel like [. . .] it can sometimes give me a lot of information and [is] super up to date. Like my TikTok feed is very political so it'll just be a lot of different stuff from different areas of life that I need to know about and I just love that.

Many participants viewed TikTok as an access point to a store of cultural knowledge. As such, maintaining a presence on TikTok led to the accumulation of social currency in their networks, and set the stage for future social interactions. As one participant explained: "mentioning a popular sound or trend gets me social currency cause everyone will be like 'oh yeah, I saw that too!' Or 'that's so funny!"

While on other social media platforms with home feeds the presentation of content often depends on users following the creators of this content, TikTok dispenses with this requirement: content is divorced from the context in which it is created. Thus, while new users of Instagram, for example, are given a moment to organize their accounts and follow the creators that they are interested in to begin curating their feed, new TikTok users are immediately presented with content: the "default state" of the platform is one of the stimulations without reflection or planning. Users need not concern themselves with anything other than consumption of content curated specifically for them, but that they themselves do not have to curate.

As mentioned, the hyper-prominence of content resulted in social interactions falling to the wayside. One participant stated that they did not use TikTok to connect with their network; the purpose of the app was purely to gain entertainment from funny videos. Other participants expressed that they would rarely follow other users or use the commenting function (except when they wanted to "work with" the algorithm to help it curate more relevant content for them). In fact, a number of participants suggested that they did not care about creators at all, and preferred to experience content on its own, estranged from its source:

> I'm using it mostly for the content. And I like the For You page because it's suggesting to me all the stuff that I like even though I might not be following the people and might not even want to actually follow them. I don't care about them beyond that video.

In addition, participants did not use the "direct messaging" function in the same manner as they would on other social media platforms (where it functions as a private messaging service). Rather, they would only use it to share TikToks with their friends, without any additional commentary. As one participant said, "the TikTok is the message." This reinforcement of the content as the main draw highlights the importance of self-making through identificatory practices in the TikTok experience. As the media text (the TikTok) is hyper-centered, the social contexts and technocultural conditions of its production fall to the wayside; rather, the significance of the media text as a circulator of meaning lies more heavily in the process of identification it facilitates in the psyche of the individual user who encounters it. In this regard, the prominence of the For You algorithm leads to the prevalence of individualized identificatory self-making processes over processes that are more overtly social and network dependent.

While interview participants were highly invested in experiencing content, they made little mention of creating it: in this regard, participant interviews departed from predictions made through analysis of the data collected with the walkthrough method. In terms of design, TikTok seems to incentivize and encourage "content creation." One of the ways in which this is done is through the design of the home screen: the size and placement of the "record" button make it more natural and easy for users to record and post videos than to find people to follow or comment on videos. Despite this leading design, the participants we talked to view themselves as observers and viewers rather than creators. This illuminates the fact that user behaviors are not always predictable, and are not entirely beholden to the suggestions made by the affordances of a platform.

Theme 3: Self-Creation Across Platforms

When asked to relate their experiences using TikTok, interview participants would often, unprompted, describe it through comparison to other prominent social media platforms, explaining how various aspects of the platform differed or aligned with features of these other platforms. Interestingly, participants seemed to lack the language to explain TikTok in its own terms, and frequently struggled to articulate where exactly TikTok fell in the larger social media landscape. Users often expressed that the TikTok experience was difficult to describe: comparisons to other platforms did not exactly "fit," or needed extensive qualifying to be accurate.

Participants most commonly compared TikTok to Instagram, Twitter, Facebook, and YouTube, often in terms of the self-making practices in which they partook. These four sites occupy distinctly different niches in the current social media landscape, and can be seen as belonging to three different "categories" according to the definitions proposed by Kaplan and Haenlein (2010, p. 62). Each category reflects a distinct purpose and offers a specific model of sociality; thus, it is somewhat curious that TikTok drew comparisons to three platforms which can be seen as differing fundamentally in ethos and organization.

Comparisons to Twitter place TikTok in the "blogging" genre of social media. This category of social media provides a stage for self-representation through the curation of content relevant to the user's identity (Kaplan & Haenlein, 2010). On such sites, processes of self and identity construction are enacted through the curation of a collection of content relevant to the user's identity. The user constructs and represents their identity through such identificatory processes, and thus, interaction types such as liking, retweeting, and reblogging are the predominant mode of engagement. Unlike other genres of social media, the central activity on such sites is not cross-user interaction through activities like commenting and liking, but rather engagement with content related to the project of self-making and from the fostering of relationships

through the act of self-disclosure. The For You page can be said to offer a version of this process of curation, as users are able to assemble a collection of media objects that represents them and with which they identify. Participants pointed out that the act of liking TikToks and experiencing content relevant to their experiences and interests was analogous to the process of scrolling through their Twitter feeds and liking or retweeting tweets that they found relatable. They also indicated that the "direct messaging" feature was used similarly on both sites: to share content they found relevant with their friends, without much additional commentary. However, TikTok differs from Twitter in a number of significant ways. Most apparent was the lack of user control over the For You feed; in contrast, the content of Twitter's home feed is almost entirely determined by the following choices of the user.

Comparisons to YouTube and Instagram align TikTok with the "content community" genre of social media sites. Such sites prioritize the sharing of media: experiencing the content is the main purpose of the platform. Thus, the aforementioned prominence of content on TikTok seems to support its placement within this category. The design of TikTok is similar to the design of content communities like Instagram and YouTube, which visually emphasize the content and feature a smaller section for comments. In addition, the function of the For You algorithm is perhaps most analogous to the algorithms used to supplement user feeds on YouTube: while users can to certain extent control what content they see by subscribing to channels, YouTube's algorithms play a significant role in determining exactly which videos a user is exposed to.

Yet, TikTok does not neatly fit into this category either. In content communities, the social aspect derives from the user following other users, and from the discussion of the presented content through comments. Such engagements with a wider network (which situate users in a community bound together by common interests) in tandem with the identificatory practices involved in selecting and curating content that can be said to represent the user, constitute the self-making processes that occur on content communities. As explained by participants, these elements are minimally relevant in the TikTok experience. Rather, TikTok engenders and supports social self-making practices beyond the confines of the digital space: through the well of culturally relevant information made available by the For You algorithm, TikTok was primarily a means through which users may gain cultural knowledge which facilitates engagement in other spaces, whether this be on other sites or in real life. Participants indicated that awareness of TikTok increased their cultural literacy and "clout" on other sites such as YouTube, as well as allowing them to "keep up" with their friends online and offline.

Finally, comparisons to Facebook place TikTok in the "social network category" of social media. Such sites prioritize communication between individuals through the activity corridors offered by the platform. On social networks, users perform their identities through interaction types like commenting, following, and friending—interactions that establish ties among members of a network. When comparing TikTok to Facebook, participants pointed out that the two platforms shared a number of functionalities: it was possible to see the activity of others in one's network, and engage with them through messages, comments, and "duets." Analysis of walkthrough method data confirmed that there were indeed a multitude of interaction types that prioritized social communication between users. However, participant interviews revealed that, while these interaction types existed, they were seldom employed by users for their intended purpose; rather, they were frequently repurposed to interact with the algorithm.

TikTok's combination of features seems to place it simultaneously in a variety of social media categories, and none at all: it can be seen as a microblogging site, a social networking site, and a content community, and yet, it departs significantly and fundamentally from each type. Correspondingly, the self-making practices it engenders (which can be broadly aligned with certain social media sites; for example, sites such as Instagram that prioritize the sharing of media may promote more indexical practices of self-making) also eschew distinct categorization.

Conclusion

At the outset of this project, we sought to understand what made TikTok special or unique in the eyes of users. Our findings revealed that TikTok differed from other social media platforms that, upon first glance, seemed to offer similar services and features. However, it did not do so through the creation of radically new features or affordances. Rather, TikTok engendered a mode of self-representation and identity creation that departs significantly from the model of "the networked self" that is found on other social media.

TikTok users occupy the precarious position of dually engaging with an external and internal entity; they engage with versions of themselves, as mediated through the algorithm. While other social media platforms facilitate interaction with other egos through a variety of methods (whether this be direct self-disclosure, the discussion of media, etc.), on TikTok the user interacts most heavily with the personalized algorithm which repeatedly confronts them with various aspects of their own personas. This model of sociality can perhaps be termed the "algorithmized self"—an extension and complication of the previously discussed "networked self"; while the latter posits that the self is created through the "reflexive process of fluid associations with social circles," the former understands the self as deriving primarily from a reflexive engagement with previous self-representations rather than with one's social connections (Papacharissi, 2011).

In addition to offering a new understanding of "the self," TikTok offers users a new "type" of social media by refusing to subscribe to an established categorization scheme. It borrows elements from a variety of pre-existing platforms, but

ultimately eludes the forms of sociality engendered by content communities, blogs, and social networks in favor of presenting a very different vision of sociality based on repeated engagement with the "algorithm."

Rather than occupying a static position in the overall internet ecosystem, TikTok's identity relative to other digital spaces is shifting; it can be said to exist on a spectrum rather than in a category. The precarious and ambiguous place that TikTok occupies in the experience of our participants indicates that current understandings of the function and purpose of social media platforms are overly simplistic. This issue extends beyond the question of TikTok's identity; the existing subdivisions—indeed, the act of categorizing in itself—fail to account for changes in usage patterns and platform functionalities that occur as platforms and users co-evolve.

Indeed, factors such as increased algorithmization have already affected fundamental changes in the operation of social media and the experiences of the users therein. While TikTok is an extreme example of the prominence of algorithms in user experience, it is an example of trends that are quickly becoming visible in other digital spaces. For example, users of Instagram report increased levels of engagement with and awareness of the algorithms that play a role in finding relevant content for their home feeds, and also engage in behaviors that aim to "work with" the algorithm to generate relevant content (such as commenting on posts to indicate to the algorithm that the content is interesting). Such practices diminish the unambiguity of Instagram's identity as a "content community."

Similar evolutions are occurring on more traditional sites such as Facebook, which once was understood as merely a platform through which users could engage in traditional social interactions such as commenting, liking, and friending. With algorithms playing an increasingly visible role in shaping the content that Facebook users see on their feeds, Facebook overlaps significantly with platforms which employ similar methods to shape user experience, such as Instagram. In this way, the division between platforms we understand as belonging in different categories collapses: the distinction between the experience of users in "content communities" and "social networks" lessens.

The ever-changing nature of digital spaces suggests that social media cannot be understood as occupying discrete categories. Nor can they be understood in a vacuum; as evidenced by the descriptions of participants who could only explain TikTok through comparisons to other platforms. Moving forward, academic analysis of social media ontologies needs to examine and interpret social media not as discrete entities, but rather as moving nodes in a more extensive ecosystem. Individual platforms do not exist separately from one another; rather, all of the elements that comprise these digital spaces interact, influence one another, and co-evolve, changing the nature and meaning of the activities and experiences of the users that inhabit them.

Author's Note

All authors have agreed to the submission and that the article is not currently being considered for publication by any other print or electronic journal.

Acknowledgements

We would like to thank Cid Decatur for his work conducting interviews through the Cornell Social Media Lab.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Aparajita Bhandari https://orcid.org/0000-0002-4844-209X

References

Anderson, K. E. (2020). Getting acquainted with social networks and apps: It is time to talk about TikTok. *Library Hi Tech News*, *37*(4), 7–12. https://doi.org/10.1108/LHTN-01-2020-0001

Anderson, M., & Jiang, J. (2018). Teens, social media & technology 2018. *Pew Research Center*, *31*(2018), 1673–1689.

Bamberg, M. (2011). Who am I? Narration and its contribution to self and identity. *Theory & Psychology*, *21*(1), 3–24.

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44.

Charmaz, K., & Belgrave, L. (2012). Qualitative interviewing and grounded theory analysis. In J. F. Gubrium, J.s A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed., pp. 347–365). SAGE.

Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, *28*(6), 164–181.

Cruz, E. G., & Thornham, H. (2015). Selfies beyond self-representation: The (theoretical) f(r)ictions of a practice. *Journal of Aesthetics & Culture*, *7*(1), Article 28073.

Frosh, P. (2015). Selfies| The gestural image: The selfie, photography theory, and kinesthetic sociability. *International Journal of Communication*, *9*, 1607–1628.

Gregory, K., Clough, P., Scannell, J., & Haber, B. (2015). The datalogical turn. In P. Vannini (Ed.), *Non-representational methodologies: Re-envisioning research* (pp. 146–164). Routledge.

Hall, S. (Ed.). (1997). The work of representation. In *Representation: Cultural representations and signifying practices* (Vol. 2, pp. 13–74). SAGE.

Hearn, A. (2010). Structuring feeling: Web 2.0, online ranking and rating, and the digital "reputation" economy. *Ephemera: Theory & Politics in Organization*, *10*, 421–438.

Hearn, A. (2017). Verified: Self-presentation, identity management, and selfhood in the age of big data. *Popular Communication*, *15*(2), 62–77.

Herrman, J. (2019, March 10). How TikTok is rewriting the world. *New York Times*. https://www.nytimes.com/2019/03/10/style/what-is-tik-tok.html

Kaplan, A. M., & Haenlein, M. (2010.). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68.

Latermedia. (2020). This is how the TikTok algorithm works. *Later Blog*. https://later.com/blog/tiktok-algorithm/

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New Media & Society*, *20*(3), 881–900.

Mirzoeff, N. (2016). *How to see the world: An introduction to images, from self-portraits to selfies, maps to movies, and more*. Hachette.

Papacharissi, Z. (Ed.). (2011). A networked self. In *A networked self: Identity, community, and culture on social network sites* (pp. 304–318). Routledge.

Papacharissi, Z. (2013). A networked self identity performance and sociability on social network sites. In F. L. F. Lee, L. Leung, J. L. Qiu, & D. S. C. Chu (Eds.), *Frontiers in new media research* (pp. 207–221). Routledge.

Rettberg, J. W. (2014). *Seeing Ourselves Through Technology: How We Use Selfies, Blogs and Wearable Devices to See and Shape Ourselves* (pp. 33–44). Palgrave Macmillan.

Thumim, N. (2012). *Self-representation and digital culture*. Springer.

Tiidenberg, K., & Whelan, A. (2017). Sick bunnies and pocket dumps: "Not-selfies" and the genre of self-representation. *Popular Communication*, *15*(2), 141–153.

TikTok Newsroom. (2021, July 01). *More Tok on the Clock: Introducing longer videos on TikTok* [Press release]. https://newsroom.tiktok.com/en-us/longer-videos

TikTok Statistics —Everything You Need to Know [Feb 2021 UPDATE]. (2021). https://wallaroomedia.com/blog/social-media/tiktok-statistics

Van Dijck, J. (2008). Digital photography: Communication, identity, memory. *Visual Communication*, *7*(1), 57–76.

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

Verstraete, G. (2016). It's about time. Disappearing images and stories in Snapchat. *Image & Narrative*, *17*(4), 104–113.

Williams, K. (2020). TikTok was installed more than 738 million times in 2019, 44% of its all-time downloads. *Sensor Tower*.

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150.

Xu, L., Yan, X., & Zhang, Z. (2019). Research on the causes of the "Tik Tok" app becoming popular and the existing problems. *Journal of Advanced Management Science*, *7*(2), 59–63.

Author Biographies

Aparajita Bhandari (MSc, Cornell University) is a PhD candidate in Communication at Cornell University. Her research interests include social media and the creation of community and critical approaches to platform and media studies.

Sara Bimo (M.A. York University) is a PhD student of Communication & Culture at York University. Her research interests include digital infrastructures, intersections between digital and physical space, and online self-making.

PX0324

Constructing Authenticity on TikTok: Social Norms and Social Support on the "Fun" Platform

KRISTEN BARTA, University of Michigan, USA
NAZANIN ANDALIBI, University of Michigan. USA

Authenticity, generally regarded as coherence between one's inner self and outward behavior, is associated with myriad social values (e.g., integrity) and beneficial outcomes, such as psychological well-being. Scholarship suggests, however, that behaving authentically online is complicated by self-presentation norms that make it difficult to present a complex self as well as encourage sharing positive emotions and facets of self and discourage sharing difficult emotions. In this paper, we position authenticity as a self-presentation norm and identify the sociomaterial factors that contribute to the learning, enactment, and enforcement of authenticity on the short-video sharing platform TikTok. We draw on interviews with 15 U.S. TikTok users to argue that normative authenticity and understanding of TikTok as a "fun" platform are mutually constitutive in supporting a "just be you" attitude on TikTok that in turn normalizes expressions of both positive and difficult emotions and experiences. We consider the social context of TikTok and use an affordance lens to identify anonymity, of oneself and one's audience; association between content and the "For You" landing page; and video modality of TikTok as factors informing authenticity as a self-presentation norm. We argue that these factors similarly contribute to TikTok's viability as a space for social support exchange and address the utility of the comments section as a site for both supportive communication and norm judgment and enforcement. We conclude by considering the limitations of authenticity as social norm and present implications for designing online spaces for social support and connection.

CCS Concepts: • **Human-centered computing → Collaborative and social computing;** *Empirical studies in collaborative and social computing*

KEYWORDS: Authenticity, social norms, affordances, social media, self-presentation, social support, disclosure

ACM Reference format:

Kristen Barta and Nazanin Andalibi. 2021. Constructing Authenticity on TikTok: Social Norms and Social Support on the "Fun" Platform. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 5, CSCW2, Article 430 (October 2021), 29 pages, https://doi.org/10.1145/3479574

1 INTRODUCTION

"To thine own self be true," spoken by Polonius in Shakespeare's *Hamlet,* is perhaps one of the better-known assertions of the value of authenticity in modern memory. The proverb underscores the importance of acting authentically, of behaving in accordance with one's own values and beliefs for avoiding self-

Author's addresses: Kristen Barta, University of Michigan, USA, krbarta@umich.edu; Nazanin Andalibi, University of Michigan, USA, andalibi@umich.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org. © Copyright is held by the owner/author(s). Publication rights licensed to ACM.
2573-0142/2021/10 – Art430… $15.00
https://doi.org/10.1145/3479574

deception and disingenuous relationships with others, and aligns authenticity with social values such as honesty and integrity. Since before Shakespeare's time, through to our contemporary moment, philosophers and social theorists have grappled with the question of self—what it is, how it is formed, and how the self exists within society. These investigations have yielded a conceptualization of authenticity as the process(es) through which an individual comes to recognize and understand one's self, including beliefs and desires, make peace with one's self, and behave in such a way that not only aligns with one's chosen self but also upholds broader social values, such as morality (cf. [45]). Prior research has also suggested that authenticity is associated with factors such as self-esteem and psychological well-being [68]. In this sense, behaving authentically in one's daily life potentially facilitates both individual and societal benefits.

As social media continues to become embedded into daily life for millions of individuals worldwide, the question of how to authentically present oneself across online and offline spaces has similarly captured scholarly attention. Scholars have noted the ways that sociality and the self become distributed across social media and other communication channels [52], [78], and investigated associations between online authenticity and well-being [20], [64]. Research has also, however, emphasized the social aspects of authenticity on social media [55], [59] and questioned the ability to enact authenticity on social media [33]; phenomenon such as social positivity bias, or the pressure to post only "highlights" of oneself and experiences [76], and other norms for self-presentation (e.g., appearing attractive or popular; [87]) complicate social media users' abilities to present themselves as complex, sometimes contradictory, and emotional beings composed of myriad identities that speak to social positions, roles, dispositions, and other facets of self.

In this paper, aligned with [33], [75], we conceptualize authenticity on social media as a socially constructed norm and explore how it is enactment on social media in a United States context; to do so we take TikTok as our site of inquiry. Used in China since 2016 under the name Douyin, TikTok exploded in popularity in the U.S. in 2020. As of August 2020, the app reported about 100 million monthly active users in the U.S., an increase of almost 800% since January. An estimated 60% of monthly users in the U.S. are between 16 and 24 years old [65, 89]; consequently, TikTok is largely shaped by and reflective of youth culture [44]. TikTok shares many features with other social media sites, such as user profiles, followers, customizable usernames, user-generated content, and interaction between users (e.g., likes, comments, and features like "duets" that allow one to duplicate and interact with others' content [15]). Yet, a defining feature of TikTok is the "For You" page as a landing site. When users open the app, they are immediately directed to a scrolling feed of content, typically from users they do not already follow, that has been algorithmically selected based on factors such as user interactions with content and accounts (e.g., liked videos, followed accounts, and posted videos), video information (e.g., sounds, hashtags, and how much of a video a user watched), and device or account settings (e.g., language, country setting) [39], [69]. Content is video-based, with most posts consisting of a video, 15-60 seconds in duration, with music, voiceover, or other "sounds" (i.e., recorded audio to accompany videos), and often conveys "goofiness" and mundanity [44].

Because TikTok is relatively new, compared to other prominent social media platforms, and because of its unique features the implications of the platform are still unfolding. Zhu et al. [90] have remarked on the potential of TikTok as a tool for disseminating public health information, and suggest that the short-video modality of the platform is effective in both providing information and in stimulating a sense of

shared emotion between the message sender and recipient. Weimann and Masri [81] note the potential of TikTok to host hate and extremism. The present study views the site's unique affordances (i.e., association between content, perceived anonymity in a highly visual context), range of content shared on the platform (i.e., from health to extremist to goofy), apparently captive audience, and TikTok's relative novelty as making it an ideal site through which to consider social norm development on social media through a sociotechnical lens.

Social norms, or "jointly negotiated rules for social behavior" [17] are formed and enforced through interaction with the members of a social group. On social media, both user behavior (e.g., likes, comments) as well as the materiality of the platform, such as policies, features and affordances, and modality of content, influence the social context from which norms arise [80]. In this paper, we use an affordance lens [18], [22] to identify sociomaterial factors that contribute to the construction, learning, and enactment of authenticity as norm on TikTok.

We draw on interviews with U.S.-based TikTok users to examine how authenticity is enacted on the site. We argue that the perception of TikTok as a "fun" space for relating goofy, quirky, and everyday experiences as well as affordances (i.e., perceived anonymity and association) of platform features (i.e., "For You" page, video modality) and policies (i.e., pseudonymous usernames, multiple accounts), contribute to perceptions of "authentic" content as normalized and valued by TikTok users. We suggest that the perceived anonymity of one's audience and self, in combination with a platform algorithm that prioritizes association between users based on proximity of interest or experience, promotes social acceptance and adoption of a "just be you" attitude that in turn supports authenticity as a self-presentation norm. We present evidence that expressions of difficult emotions and experiences, as well as positive ones, are socially accepted on TikTok, and posit that displays of emotional "rawness" fall within the bounds of normative authenticity. We then provide insights on the comments section as a site of norm validation and judgment. We argue that authenticity on TikTok appears to describe both "unfiltered" content and more selective presentation of a *partial* but *still* authentic self. We conclude by reflecting on the implications of authenticity as norm on TikTok for designing sociotechnical spaces to encourage emotional expression and social support exchange. These interactions between social and material factors carry implications for designing supportive and compassionate social media spaces; we reflect on some of these possibilities and identify areas for further research. Despite the potential benefits of normative authenticity on social media, we are cognizant of the likely limitations of these benefits; artifacts such as social media are inherently political [83] and can perpetuate marginalization of already marginalized identities, as has also been observed on TikTok [12], [43], [56]. In this exploratory paper, we may only speculate, but conclude with discussion of the limits and unintended consequences of normative authenticity on social media.

2 LITERATURE REVIEW

To situate authenticity as a social norm that influences emotional expression on social media, we first define social norms in the context of self-presentation and connect this definition to current understandings of authenticity on social media. We then present affordances as a lens for understanding the sociomaterial influences on norm development and perpetuation and review the literature on association and anonymity as potentially salient to authenticity on TikTok. We then connect these social and material factors to speculate as to the potential effects of normative authenticity on TikTok as a space for social support.

2.1 Social norms

A norm refers to a pattern of behavior, acknowledged by members of a group, that "bounds acceptable behavior" [40]. Norms are socially constructed [62], which is to say they are informed by other standards of conduct, such as codified rules, social values, and material constraints [88] and collectively established by a group's members. On social media, factors such as privacy settings, how users connect (e.g., following mechanisms), and modality of content (e.g., video, text) inform the creation of platform norms [80], including self-presentation and emotional expression norms. Norms may vary across platforms [57], [88] due to differing configurations of user demographics, use motivations, and available features [75]. Individuals may observe interactions between other network members to discern norms prior to becoming active on a platform [3], [13], [46]. For example, when deciding how to engage with social media posts about difficult emotions or stigmatizing experiences, people tend to observe and assess others' engagements with the posts [2]. Observing norms in action may be advantageous, as self-presentation and interaction norms are socially enforced; norm violation may be met with sanctions such as confrontation from an/other platform user(s) or loss of use privileges [57].

Broadly, social media norms for self-presentation often involve presenting oneself in a positive light, contributing to a perceived social positivity bias in which positive self-expression and content is more common and receives more interaction from others than negative self-expressions [24], [64], [76]. For instance, Yau and Reich [87] found that appearing interesting, likeable, and attractive were powerful self-presentation norms for adolescent social media users. As such, more specific norms, informed by user characteristics and other factors, may contribute to, uphold, and reinforce a positivity bias on a given platform.

An outcome of positivity bias is a perception, reinforced through affirmation and validation (or lack thereof) from other social media users, that expressions of negative or difficult emotions are inappropriate for social media [37]. In particular, negative self-disclosures may be deemed inappropriate for public, undirected (i.e., not targeted at anyone in particular) communication on social media [53], and less visible channels, such as private messages, may be deemed more appropriate for sharing difficult and intense emotions [11]. For example, individuals experiencing distress and stigma find platforms like Facebook, where they are connected to people they know (not strangers) and identified by their names, as a site on which sharing difficult emotional experiences is outside the norm [1]. Certain platforms may also be perceived as more appropriate spaces for difficult emotions than others [5], [54]. In a survey of Dutch social media users, for instance, Waterloo et al. [80] suggest that users perceive negative expressions as more appropriate on platforms like WhatsApp and Twitter than on platforms like Facebook and Instagram. Despite variance across platforms, the presence of positivity bias across social media spaces requires users to implement strategies for expressing difficult emotions.

When context determines the propriety of feelings, expressions of divergent feelings must be altered to "fit" the context [42]. For instance, the popularity of "finstas" (a portmanteau of "Insta" and "fake"), secondary Instagram accounts often reserved for smaller, more selective/deliberate audiences and less polished content [77], [88], may be indicative of the pressure to present positively and offer one strategy for circumventing this perceived expectation. The use of humor as a form of expression offers another strategy for managing the appropriateness of negative expressions. That is, humor may "mitigate the gravity of negative emotions" [86] and make negative expressions both less likely to be sanctioned and more palatable to other users. Humorous content, such as memes, may also provide a means of expression

on stigmatized topics, such as mental health [7], [36]. Through similar mechanisms, humor may undercut a pressure to present a curated self and allow a more socially acceptable way of authentically declaring "look I'm a real person, my life is a mess too" [19]. Thus, how people present difficult emotions, via humor or other strategies, in addition to where (i.e., platform), may affect whether such emotional expressions are perceived by others as authentic. This background and the presence of humous content on TikTok informed our attention to these aspects in the analysis we present in this paper.

In the following sections, we define authenticity before considering the material affordances, such as discursive anonymity, and features, such as modality, of TikTok that, as we will show in our analysis, further shape authenticity as a self-presentational norm.

2.2 Authenticity

Scholars trace authenticity's conceptual roots back to philosophers such as Socrates [45], though its modern origins are often credited to Heidegger and existentialism [34]. Such philosophical orientations position authenticity as at once individually and socially determined, and as ultimately affirming social values (e.g., integrity, honesty, morality). Simply stated, authenticity may be defined as "unobstructed operation" of one's true self [45], or "accurate" self-presentation [10], yet these succinct encapsulations smooth over myriad dimensions and sociality influential to understanding authenticity in online contexts. "Self" has also been variously defined (cf. [50]), though here we understand it to mean how one thinks of oneself. We draw on an adjacent construct, self-concept, to position self as referring to the many identities, roles, beliefs, and values that compose an individual (e.g., I am a woman, I am a teacher, I am kind; [49]). What is deemed authentic (and how it is judged) is informed by the identity or aspect of self in question [38], [58].

Psychologists view authenticity as involving interactions between one's experience, self-awareness, and behaviors [8], [84]. Agreement across these dimensions (e.g., behaving in accordance with one's beliefs and values) is thought to increase feelings of authenticity. Understood as such, authenticity may correlate with outcomes such as psychological well-being and self-esteem [20], [64]. However, Jongman-Sereno and Leary [41] have questioned the relationship between authenticity and well-being and argued that inconsistent conceptualizations of authenticity may lead to operationalizations that confound authenticity with similar variables, such as honesty. Hardt [34] also notes that behaving authentically, or in accordance with one's own intrinsic beliefs, requires freedom and may mean acting against dominant social mores. There may thus be limits to authenticity when positioned as unilaterally beneficial, as one may encounter tension between acting as one's true self and the limits of social propriety. Self-assessment of authenticity may be complicated by the fact that knowing oneself (i.e., self-awareness) and identifying intrinsic beliefs may be rather difficult [82], or even painful [45]. Thus, understood as a psychological term, whether authenticity is always desirable or even discernable is debatable [41].

In online contexts, authenticity is similarly conceptualized as multidimensional, but is likened to *performing* authenticity. As Gaden and Dumitrica [26] observe, on social media, "the 'authentic' self appears to be a process of witty narration of an inner personality made available for subsequent consumption by the audience." Gilpin, Palazzolo, and Brody [27] similarly emphasize performance and define authenticity in the context of online political engagement as a performance in which one appears credible (authority), reliable and genuine (identity), forthcoming (transparency), and open to interaction with others such as political constituents (engagement). Understood as performance, the outcomes associated with online authenticity may differ from those identified in psychological perspectives.

Reinecke and Trepte [64], for instance, have questioned the association between authentic self-presentation on social media and well-being, given the aforementioned social positivity bias observable in many online spaces, and related associations, such as that between social comparison and self-esteem (e.g., [79]). In other words, authentic self-presentation on social media may benefit the well-being of those who already experience higher levels of well-being or self-esteem, as they have less of a need to express difficult emotions [64]. While social comparison certainly exists offline (cf. [23]), online spaces may concentrate and amplify expressions of personal success in a way that exacerbates both perceived social positivity and maladaptive comparison.

Authenticity on social media may thus be better understood as an attribute of self-presentation, which, like norms, is shaped through interaction with the affordances and audience(s) of a particular platform [33], [55]. Indeed, previous works emphasize that online authenticity is "dependent on subjective evaluation by participants or observers" [27] and that the ability to create and determine authenticity "does not only reside in encoders or creators, but also in decoders or interpreters" [59], underscoring the role of audience and audience responses (e.g., comments) in determining authenticity [55]. It thus follows that, as authenticity itself is socially constructed, authenticity may also be informed (i.e., amplified or constrained) by other self-presentation norms of a given social media platform. In a comparison of Facebook and Last.fm, for instance, Uski and Lampinen [75] note that "being real," a value shared by users in both spaces, was bounded and qualified by other self-presentation norms, such that "being real" required not oversharing and not seeking attention through profile updates. As previously mentioned, TikTok's user base is reputed to be quite young, and content on TikTok is often understood as "goofy" [44]. These social factors and perceptions have the potential to shape norms for self-presentation on TikTok, including authenticity. We explore these factors in our eventual analysis.

We thus consider authenticity as aligned with scholars such as Uski and Lampinen [75] and Haimson and Hoffmann [33], the latter of whom position authenticity as a socially constructed, artificial category, in which effortless yet sincere presentations of self that conform to the expectations of an audience as well as the context of expression (i.e., social media) are read as authentic by onlookers (e.g., networked others). Given the socially constructed nature of both authenticity and social norms, we ask:

> RQ1: How is the self-presentation norm of authenticity learned, enacted, and enforced by users on TikTok?

2.3 Affordances

We use an affordance lens to better understand the role of platform materiality in shaping authenticity as a TikTok norm. We define affordances as the abilities that arise from interactions between users, their goals, and features of social media, and which enable or constrain behavior [22]. Affordances are not social media features themselves, nor behavioral outcomes, but rather abilities of a technology (or social media feature) that an actor perceives as relevant to their behavioral goals (e.g., a rock may be used as a hammer or as a paperweight, depending on the situation). Affordances are also variable [22] and graduated, such that technologies, as artifacts, may demand as well as allow, discourage as well as refuse, a behavior [18]. Treem and Leonardi [74] argue that persistence, editability, visibility, and association may be especially pertinent to facilitating communication among work team members; these affordances may vary depending on factors such as an actor's perception of a technology's abilities, the actor's communication

needs, and cultural (or institutional) norms surrounding technology use [18]. As such, like norms, affordances are dependent on context.

In examining TikTok's affordances and their role in shaping norms of authenticity, we explore relationships between affordances as well as the outcomes affordances facilitate, such as emotional expression and social support. Given our framing of authenticity as a social norm, in which norms are socially constructed, as well as our interest in emotional expression and social support, we emphasize affordances that directly affect social interaction: association and anonymity. That is, while other affordances certainly have social dimensions (the permanence and searchability of content may adversely affect a friendship, for example), we view association and anonymity as directly related to what one shares about oneself and with whom it is shared, which has implications for authenticity.

2.3.1 Association. Association, defined as connections between users, users and content, or content and content [74], describes one way that user networks become solidified on social media. Association between users varies by platform, though can generally be thought of as either reciprocated (as on Facebook, where users must mutually agree to be connected as "friends") or unreciprocated (as with followers on Twitter) by design. Features such as follower lists can also indicate connections between users [21]. Most platforms afford association between content and users, in that content often bears the creator's username or can be found otherwise connected to a user's profile/account. Hashtags on Twitter afford association between pieces of content, in that tagged content (e.g., a tweet tagged #MeToo) will appear in search results, thus associating content with that hashtag and suggesting that similarly tagged posts may be topically related.

TikTok similarly affords association between users, users and content, and content and content. Unlike other platforms, however, association between users may not be the primary mechanism for delivering content. For example, on Facebook, one's "feed" is primarily content shared by established "friends," other users to whom one is reciprocally connected. On TikTok, association between users as a means for providing content appears secondary to algorithmic determination of association between content and content. That is, TikTok's "For You" landing page draws on user data to suggest content that is, ostensibly, thematically proximate to other content liked or interacted with by that user [39]. In other words, it is not necessary to establish connections with other users on TikTok in order to receive content from other users. While building a network is certainly an option (opposite the "For You" page is a "Following" page), the "For You" page as default landing page encourages users to interact with content and build networks based on affinity and similarity of content, rather than promoting users who may be connected to others in one's network (e.g., suggested connections). In this way, association afforded by TikTok appears to function differently from many other social media spaces and may affect the way that self-presentation norms are developed and circulated on the platform—an aspect we explore in this study. Further, as we address in 2.4, encouraging connection on the basis of similarity may have implications for social support processes.

2.3.2 Anonymity. In communication theory, anonymity is understood in terms of awareness of the source of a message and may be defined as the degree to which a message source is perceived to be unknown [6]. Technology scholars, however, are quick to note that perceived anonymity (i.e., feeling as if one is unidentified/fiable) differs from "true" anonymity, as trace data such as IP addresses and other information collected by internet service systems makes identification possible [25]. Anonymity is thus variable and dependent upon contextual factors such as audience (e.g., the author of a comment may be anonymous to another forum user, but not to a forum moderator). Anonymity is multidimensional, comprised of discursive anonymity and visual anonymity [67]. Discursive anonymity aligns with

Anonymous' [6] definition of anonymity as lacking attribution to a source, whereas visual anonymity refers to a lack of visual representation, such as through photographs [63]. These forms of anonymity are also variable, as certain identity cues (e.g., legal name) may be more "identifying" in isolation than others (e.g., location) [6], [63].

Similar to Instagram, TikTok emphasizes visuality by virtue of its content modality; to create a post on Instagram, a user must include an image, and to create a TikTok, a user must include video of some kind. This emphasis on visuality makes maintaining visual anonymity on TikTok more complex than on other platforms (e.g., Reddit which is primarily text-based and pseudonymous). While it is indeed possible for a user to post videos that do not disclose their visual identity, many users do present themselves in content. Consequently, the visual aspect of TikTok complicates what it means to be anonymous on the platform, and challenges expectations informed by anonymity, such as disinhibited communication and emotional expression [5], [72].

Disinhibited communication is a potential outcome of anonymity on social media. The disinhibition effect describes the behavioral phenomenon of people speaking and acting in online spaces in ways that they ordinarily would not in face-to-face contexts [72]. A mechanism informing this effect is reduced risk associated with anonymous communication [5], [20]; that is, disinhibited but identified communication may result in consequences (e.g., employment termination) in a way not possible with anonymous communication. Disinhibition supports both beneficial behaviors, such as disclosure and social support exchange [5], and detrimental behaviors, such as harassment and trolling [72]. Discursive and visual anonymity, as factors contributing to disinhibition [69], may be especially relevant in computer-mediated communication contexts [37]. Indeed, TikTok poses an intriguing case regarding these dimensions, as the platform affords discursive anonymity through features and policies that automatically-generate random usernames on account creation and allow customizable, pseudonymous usernames, as well as multiple accounts, yet is highly visual in that the modality of content is almost exclusively short video. Visual anonymity may also be asymmetrical on TikTok, in that video creators and commenters may have differing levels of discursive and visual anonymity, which potentially complicates how disinhibition manifests on the platform.

Scholarship that explores the effects of visual anonymity on disclosure and disinhibition is growing. In an analysis of blog content, Hollenbaugh and Everett [37] found that visual identification of bloggers, such as a representative photograph, was positively associated with disclosure, meaning that the more identifiable one was, the more information one shared on the blog. This seemingly contradicts the predicted effects of online disinhibition [72], though the authors suggest that "visual cues may go beyond simply identifying someone to instead constituting an important component of self-disclosure overall" [37], and call for further scrutiny of visual anonymity as contributing to disinhibition. Visual presence may also afford a different range of cues through which to self-disclose or signal identity without compromising discursive anonymity. This may be especially impactful in the context of enacting authenticity on TikTok, as video and photographs ostensibly provide visual proof of events and emotions [32], [70].

In sections 2.3–2.3.2 above, we have connected association and anonymity as affordances to features and policies of TikTok, such as the "For You" page, pseudonymous usernames, and video modality. These material factors contribute to TikTok's social context, which informs social norm development, but in and of themselves do not determine norms. We thus ask:

RQ2: How do the material features and affordances of TikTok inform authenticity degrees as a norm on TikTok?

2.4 Social Support and Affordances

As affordances are variable and dependent on context, the outcomes supported by association and anonymity are similarly multiple. In this section, we define supportive communication before highlighting how association between content and discursive anonymity may potentially facilitate social support solicitation on social media, in addition to normative authenticity and difficult emotional expression, as previously discussed. We define social support broadly as "the things people say and do for one another" [29], and use supportive communication as a lens through which to understand such actions. Supportive communication emphasizes verbal and nonverbal messages intended to communicate assistance to others in need of aid [14], and as such is applicable to an online communication context. In addition to comments, paralinguistic digital affordances (PDAs), such as likes and favorites, may communicate support or validation [16], [66]; consequently, comments and other user interactions on TikTok may be impactful for both communicating support as well as for validating or sanctioning content as adhering to/deviating from normative authenticity.

Like acceptable emotional expression, the efficacy or amount of support PDAs communicate may differ across platforms due to factors such as audience and support provider [35]. Indeed, "significant others" (as used by Thoits [73], similar to Granovetter's [30] strong ties) and experientially similar others may offer different forms of support that may be efficacious in differing contexts; broadly speaking, significant others may facilitate a sense of importance (or "mattering" to someone) and self-worth, while experientially similar others may provide empathic understanding and validation of experience/reaction [73]. In a formal support context with experientially similar others, such as online Al-Anon meetings, mechanisms underlying supportive communication may include fostering a sense of belonging in seeing similarity in others' experiences, reconstructing one's self-concept, and contributing resources or other steps for recovery [47]. Relatedly, individuals seeking support on social media tend to have unique expectations about receiving support from significant others compared to anonymous sympathetic strangers, where anticipations of received support shape support seeking decisions [1]. Such findings suggest that association between users based on similarity of interests or experience, as is the case on TikTok, might facilitate support exchange, but that remains to be known. Additionally, anonymity afforded by online spaces may facilitate both disclosure of sensitive information, as previously discussed, and supportive responses to disclosure, in a way that does not necessarily amplify negative disinhibition such as harassment or trolling [5].

The above reviewed literature illustrates the theoretical connections among affordances (such as association and perceived anonymity), normative authenticity, emotional expression, and social support. Given the potential for affordances such as anonymity and association to facilitate multiple outcomes, such as emotional expression and social support, we ask, as our final RQ:

RQ3: What does the construction of authenticity on TikTok teach us about how to design social media spaces to facilitate the sharing of both positive and difficult emotions for social support solicitation and provision purposes?

3 METHODS

We conducted semi-structured interviews (N=15) with self-identified frequent TikTok users. A recruitment service was used to identify potential participants and administer a screening survey. The survey was used to identify individuals who met minimum eligibility requirements: have used the TikTok app for at least six months, use TikTok at least once a day, live in the U.S., and be at least 18 years of age at the time of the survey. A total of 284 responses were recorded, of which 257 met the eligibility criteria. We invited 27 survey respondents to participate in interviews and completed interviews with 15 individuals. Participants were selected and invited based on their responses to survey questions; we prioritized participants whose responses we felt suggested the potential for rich data. For example, participants whose survey responses included details and variety in terms of experiences using TikTok were prioritized over those with vague responses. Additionally, we purposefully selected participants to provide a range of perspectives along dimensions of race, gender, sexuality, and age in the final sample. As is common in qualitative research, saturation of themes informed the final sample size [28]. We obtained informed consent from all interview participants. Participants were offered $20 gift cards as compensation. The university IRB approved the study design.

In interviews, we asked participants about their TikTok usage habits, including production and consumption of content; understanding and conceptualization of the app; navigating the functionalities of the app, including how they sought out content; and perceptions of the algorithm and recommendation mechanisms of TikTok. We conducted all interviews via Zoom's video/audio calling services; we invited participants to have the app open during interviews to more accurately relate their experiences and how they made sense of them or to provide examples where possible. We recorded and transcribed all interviews. Interviews lasted an average of 75 minutes, ranging from 48 to 107 minutes, depending on participant responses.

Table 1. List of study participants.

Participant	Age	Gender	Sexuality	Race/Ethnicity	Education
P1	21	Female	Bisexual	Black	Some College
P2	23	Female	Straight	Black	Some College
P3	36	Female	Bisexual	White	Some College
P4	44	Female	Heterosexual	Black	Undergraduate
P5	42	Male	Straight	White	Postgraduate
P6	26	Male	Gay/Queer	Hispanic/Latino	Some College
P7	19	Female	Straight	White	Some College
P8	50	Female	Straight/Asexual	White	Postgraduate
P9	45	Female	Heterosexual	Black	Postgraduate
P10	18	Male	Heterosexual	Asian	High School
P11	18	Female	Straight	Indian	Some College
P12	20	Female	Straight	Black	Some College
P13	28	Male	Gay	Black	Undergraduate
P14	18	Female	Asexual	Multiracial Asian	Some High School
P15	21	Female	Straight	Asian	Some College

We coded all interview transcriptions using Dedoose, a qualitative coding software. The second author and two research assistants (RA) (involved in the larger project of which this study is a part) used line-by-

line open coding to establish initial codes; the RAs and the second author independently open-coded one transcript [71] and compared resultant codes, leading to refined codes through detailed discussions. Following this coding check, one RA coded the rest of the data, and the second author discussed developing codes weekly with them and iteratively and collaboratively refined the codes. The first author then reviewed all the interviews, confirmed coding, and used memoing [51] to identify connections across codes. That is, based on the confirmed codes and memos, the first author developed themes that focused more explicitly on platform utility, features, affordances, and audience/network structure. Emotional rawness, authenticity as valued by users, and a "just be you" attitude emerged as themes in codes regarding perceptions of TikTok and comparisons to other platforms, addressing RQ1. Further analysis of codes regarding specific features of TikTok, such as the "For You" page and network structure, surfaced affordances that appeared influential to users' experiences with emotionality and authenticity, and addressed RQ2. The first and second author discussed discovered themes during this process on a weekly basis and refined the themes and connections between them by going back to the data iteratively as needed. Discussing the potential outcomes of sociotechnical affordances addressed RQ3.

4 FINDINGS

Participants reflected on myriad attitudes regarding behaviors exhibited on TikTok. In the following sections, we draw on these insights to suggest that particular features and affordances of the platform—the "For You" page, commenting mechanism, association between content, and perceived anonymity—contribute to and uphold authenticity as a self-presentation norm on TikTok. We emphasize association and anonymity as affordances of interest given the aforementioned implications for sociality, norm development and judgment, disinhibited communication and expression of difficult emotions, and social support. We suggest that normative authenticity becomes further apparent through consideration of TikTok as the "fun" platform, as well as a "just be you" attitude and emotional rawness as valued qualities of content. As normative authenticity is socially constructed and enforced, we also show how user comments can be a site of norm judgment and sanctioning, as well as a potential site of social support provision.

In responding to RQ1, we first consider participants' overall perceptions of TikTok. We find agreement with others who describe TikTok as "fun" and as supporting "goofy" or mundane content [44]. More specifically, we identify a "just be you" attitude that appears to apply generally to content on TikTok. As social and material factors are mutually influential in norm development, we then present findings in response to RQ2 to consider participants' perceptions of audience and network on TikTok, and show how perceived anonymity—discursive (but not necessarily visual) anonymity of oneself and of one's networked others—as well as association between content uphold this "just be you" attitude and further construct authenticity as a self-presentation norm. We then show how these factors interact to support the expression of difficult emotions on TikTok, and how these expressions also fall under the umbrella of normative authenticity. We also consider the comments section as a feature of TikTok influential in norm judgment and enforcement and extend this discussion to respond to RQ3.

4.1 Perceptions of TikTok as "fun"

In responding to RQ1, we focus on user perceptions of TikTok as a space for "fun" content. TikTok describes its mission as "to inspire creativity and bring joy" [60]. As P8 explained, "[TikTok] it's just completely different because I feel like it's more for fun." Perceptions of platform utility, as suggested by this comment,

contribute to the range of behaviors deemed appropriate for a social media space. Perceptions of TikTok as "fun" also coincided with perceptions of the app as facilitating freedom of personal expression. P3 explained,

> "I would say that it's [TikTok] a place where you can be silly. You can be funny. If you want to be a different person, you can be a different person. If you want to be yourself, you can be yourself."

In the context of this conversation, being a "different" person was linked to costume play (cosplay), which again suggests room for identity play and sharing hobbies and interests on TikTok. In this case, being a different person does not necessarily equate to inauthentic self-presentation; rather, a user might communicate an authentic interest in cosplay by sharing their work.

Like other platform values, "fun" is perceived, learned, and internalized by other users [62]. P1 touched on the process for learning site norms by explaining that observing others, "seeing other people having fun in their own way and making it their own," led them to the realization that "I should just be me, that's what's fun. That's what people want to see." Others similarly remarked on a learning curve for the app and arriving at more personal content through trial and error. P6 reflected,

> "I guess I wasn't getting the flow, so to speak. So I was trying to make funny things that weren't really funny. And then I just switched over to […] I started making videos just of me talking, sharing my really embarrassing, funny hookup stories and such."

This example underscores the importance of authentic self-expression to the "fun" value on TikTok. That is, P6 found greater success in sharing personal stories with humor than in imitating others' content or pursuing humor that wasn't naturalistic or otherwise in harmony with their personality and identity.

Related to the overall perception of TikTok as a "fun" platform, an attitude of "just be you" also appeared prominent. This attitude manifested in part as disinterest in participating in TikTok trends, as some users viewed following social trends as inauthentic or otherwise discordant with their understanding of self. P12, for example, explained that they do not recreate trends in their own videos, "because it's just not what my page is about. […] I don't mind watching other people and I'll even interact with liking and comment[ing], but I don't really want to post that. That's not me." P10 similarly expressed an appreciation for originality as connected to authenticity, noting,

> "I want to be a content creator, not a copier, right? […] I feel like that's the artist in me being like, 'Yo, you got to do your own thing. You got to be yourself.' I don't just want to copy someone else's audience. I want my own."

In invoking others' audience, P10 may be alluding to a perceived consequence of TikTok's emphasis on association between content (discussed further in the next section), in that individuals who post similar content or about similar topics may attract similar audiences. Thus, being an authentic individual on TikTok may also extend to one's audience, in that gathering an audience perceived to be "my own" reflects successful enactment of "being myself."

That said, as on other social media, the pressure to self-present an ideal version of oneself exists on TikTok. However, we find that this ideal self may still be linked to and informed by a value for authenticity. P9, for example, explained that they made TikToks depending on how they felt that day: "If I'm happy, I do one. I don't do one when I'm sad or when I'm depressed. […] Because I want people to see me as being happy, not see me as being depressed. […] Because that's me." In this way, showing oneself being happy

was still understood as being authentic or true to oneself, albeit a partial representation of a more complex authentic being. This example aids in underscoring the connection between authenticity on social media and self-presentation by demonstrating that authenticity can be selective—authentically presenting socially desirable facets of oneself does not necessarily require similarly presenting facets perceived as less socially acceptable or undesirable.

Social media users seem to be aware of positivity biases on platforms and the pressure for users to curate accounts that display only "highlights," rather than emotional complexity. This awareness of skewed self-presentation may similarly underlie perceptions of authenticity on social media. As P10 explained,

> "I don't know, to me, it's like, if you're going to try to go viral or be a content creator, at least be honest with yourself about it. …People want to go viral. People want attention. That's just what social media is for."

This assessment of social media's broader utility lends complexity to our understanding of authenticity on social media. More specifically, it suggests that, as a norm, authenticity is socially constructed and promoted as valuable, yet individuals retain some agency in determining what their authentic self looks like. That is, some users may find authenticity, as a norm on TikTok, to be freeing and resultant in disinhibited self-presentation in a way that feels authentic, while some users may find authenticity as valued on TikTok to be simply another lens through which to filter self-presentation, as evident in posting only content that authentically shows one as happy. In other words, authentic content on TikTok is not necessarily synonymous with unfiltered content, though we find that there is room for both on TikTok.

4.2 Authenticity as emotional "rawness"

As the previous section suggests, the attitude of "just be you" accompanies both "fun" in humor, dance trends, and other forms of creative expression, as well as more intimate expressions, such as relating dating experiences. In another sense, "just be you" also manifests in emotional "rawness" or expression of difficult emotions/emotional experiences. P11 referred to reactions they'd seen to TikTok *on* TikTok, explaining, "TikTok, [creators] say like, 'Y'all have no filter,' or, 'Y'all are too comfortable on this app.'" They continued,

> "They're telling details, really intimate details or just really personal things that had happened to them. […] That's what our generation does. They make jokes about important events and important life changing things that happen in their life, whereas on Instagram, people are a lot less personal. It'll be just a cute picture of them somewhere and that's it, but no caption, no nothing."

As this example suggests, the emotionality of TikTok is perceived as heightened in comparison to other social media spaces, such as Instagram. Additionally, P11 links humor on TikTok to personal disclosures regarding momentous (and not necessarily beneficial) events and experiences. As discussed, humor may be a tactic for making otherwise difficult emotional expressions more palatable to others and less likely to violate norms of social positivity and expression. In this instance, humor is linked to generational affect more broadly, such that humor may not be employed to soften negative expressions, but rather to further communicate authenticity through humor as a signal of generational belonging.

Participants considered the ability to openly express emotion, particularly difficult emotion, as fairly unique to TikTok as well as an outcome of perceived anonymity on the platform (discussed further in relation to RQ2). P3, for example, expressed surprise at this quality of TikTok, saying,

> "Just the ability they have to open themselves up and just share with people that they don't know, just their raw emotion. … It's surprising because sometimes I have trouble doing that, and they seem to be able to do it."

P4 also distanced themself from the practice of opening up on TikTok, but drew a connection between freedom of emotional expression and relative lack of consequence, explaining,

> "I think it [TikTok] makes them feel more comfortable, because they were like, how can they judge me? I'm just showing raw emotion, the only thing they can do is send me a bad comment, but I don't have to check those comments."

As both of these comments imply, "raw emotion" is apparent and normalized on TikTok; consequences of expressing difficult or raw emotion appear to be tempered by qualities of the app, such as an audience perceived and felt to be anonymous and user control over the visibility of comments. We expand on these factors in the next section.

More broadly, the ability to express emotion without or with less fear of judgment may facilitate more intimate disclosures as well as more mundane communication than generally supported on other platforms. In illustrating the former, P12 commented,

> "Now, there's no fear anymore. So, people post freely and just are frustrated and they want to get the word out. So, I feel like, especially on TikTok, they post a lot more about problems they've been facing for who knows how long. Now, TikTok is a[n] outlet for them to speak up about it and get the word out."

In the context of this comment, P12 refers to social issues, such as police brutality and violence perpetrated against Black communities in the U.S. In this way, TikTok may facilitate disclosure of negative emotions motivated by desires to vent about issues affecting oneself or one's community as well as to raise awareness or educate others [2].

In another sense, the acceptability or lower stakes associated with disclosure may also support relatively mundane disclosures. For example, P6 noted, "TikTok is [for when] there was a moment in my day where I just wanted to share this with you." Similarly, P3 reported sharing details of a recent hospital stay on TikTok:

> "I told them what was going on. Well to a point. Some of it was TMI. But I just kind of was like, 'If you guys want to talk, message me. I'll try to do some videos,' which I did. I tried to kind of give them a sneak peek into what is going on just in a normal person's life."

While a hospital stay may not be a "mundane" activity for many, and may in fact be quite stressful, P3's emphasis on sharing "a normal person's life" and P6's desire to share moments in their day provides support for authenticity as a norm of expression on TikTok, regardless of how novel or exciting the event precipitating that sharing may be.

4.3 Constructing authenticity: Affordances and features of TikTok

In responding to RQ2, we found that participants' perceptions of TikTok, particularly as they pertain to platform functionality (e.g., information seeking, social connection, entertainment) and audience, contribute to the salience of authenticity as a norm. Function and audience perceptions also aid in illuminating how TikTok differs from other social media spaces. Our findings highlight anonymity, as both

audience perception and affordance, and association between content as factors further contributing to TikTok's reputation as a "fun" social media site.

4.3.1 Anonymity to an audience. Participants' networks on TikTok varied, but many indicated that friends and known others were not part of their network or audience. Responses varied as to whether and to what degree audience composition was determinable by users, and to what extent this audience was simply an outcome of the platform's structure. As P3 explained, "On TikTok, [my audience] it's people I don't know. It's strangers." P1 similarly characterized TikTok broadly as "just a bunch of strangers." Others noted a convergence between the difficulty of finding known others and a disinterest in connecting with them on TikTok. "I wasn't able to find my friends on here, and I don't plan on doing that," P11 explained. They continued, "I guess it is different [from other social media] because it's harder to find people." The prominence of strangers as a perceived audience contributed to a sense that TikTok was more anonymous than other platforms, where participants were connected to family, co-workers, and/or friends—in other words people with whom they had pre-existing ties.

Participants also alluded to profile information and audience size as potentially contributing to perceptions of anonymity, though these were less commonly explored features. P8 suggested that users provide relatively little personal information on their profiles, explaining, "I don't think they can learn much about you on TikTok. It's really a hard way to do it because I don't think you really give a lot." This comment intimates the influence of discursive anonymity on broader perceptions of anonymity; information disclosed through content (e.g., visual identity, physical location or surroundings) was not a factor in this participant's assessment of anonymity on TikTok. Connections between anonymity and disinhibition were also apparent. P10, for instance, noted the disinhibiting effect of having a small audience, saying, "I don't have big enough of a following to worry about security yet. So I'm chilling." Here, "chilling" implies a lack of perceived audience risk associated with posting content. These comments suggest that anonymity on TikTok is informed both by audience composition and size; the ability to have an audience of previously unknown others that is also small may contribute to a sense that one is also anonymous, which in turn may reduce perceived risks associated with self-expression and encourage disinhibited expression.

4.3.2 Anonymity and association. Perceptions of anonymity were further supported by a perceived lack of association between accounts; features like the "For You" page as a default landing page upon opening the app and limited visibility of association between TikTok users contributed to this overall sense of anonymity. Association, or ties between users, users and content, and content and content [74], on TikTok was often explained through comparisons to other social media platforms. For example, P4 compared association on Facebook to TikTok, saying, "on Facebook you have to be friends with that person, to see that emotion or to see what's going on. On TikTok as soon as you pull it up, there's people you ain't even following." In other words, Facebook's reciprocated association stands in stark contrast to the ability (by design) on TikTok to view content from non-networked others. Similarly, P8 characterized Facebook as

> "People stay on there to stay connected. TikTok is, there's no connection. You can just look at videos all day. You can follow someone, but that's the extent of your knowledge of them. They could be anyone. You're just amused."

P8 thus also alludes to the interaction between association and anonymity in defining the functionality of TikTok. It is not a space for connection, like Facebook; rather, it's a space for amusement and that is all one "knows." As these examples make clear, we find that TikTok "feels" different from other social media spaces in the sense that content may be consumed without much awareness or consideration of users or even

one's network in the space. As such, anonymity of both oneself (i.e., discursive anonymity) and one's network may contribute to both disinhibited communication as well as a network based on experiential similarity rather than extant relationships (i.e., significant others).

The emphasis on association between content, in that the "For You" page is algorithmically curated based on a user's previous interactions with content, rather than association between users, also supports a perception of TikTok's utility as almost a break from one's pre-existing and other social networks. P7 explained,

> "[On TikTok] it's less looking at what my friends are posting and more just looking at things that people post in general that are interesting to me and that I enjoy. Whereas on other social media, I feel like I'm just focusing on my friends and what they're doing."

As such, TikTok content can feel more personal or personally relevant, regardless of connection to the creator. P7's comment also suggests that there may be a social obligation to keep track of information about friends and engage with their content. That connections to creators are not emphasized by TikTok's structure (though following others is possible and a "Following" feed is available opposite the "For You" feed) appears to instead emphasize one's connection to content.

Some participants explained that this emphasis on content being thematically proximate to a user's interests can result in different or more meaningful feelings of connection between viewers and content, or spark recollection or reflection on one's personal experiences. For example, P9 expressed appreciation for content about grandmothers: "Just the love that people have for their grandmother. And how much I love my grandmother, how much I cherish her. Things that they talk about with their grandmother moves me." Although this content is not unique to TikTok (as the participant went on to explain), the availability of such content—through the "For You" algorithm—may be, and was notable for the participant. Similarly, P7 related a connection to content posted by a former American Girl Doll store employee:

> "I used to go to the American Girl Dolls store and salon all the time to get my doll's hair done. I just connected to that because it was a childhood memory that I don't really think about often, but it was interesting to see."

While the identity of the creators behind the content may not be memorable, the emotional responses to content that stem from one's memories and personal experiences may be impactful for TikTok users. TikTok's emphasis on association between content makes these moments more available to users and as such shapes users' perceptions of the platform overall.

The prominence of association between content on TikTok also interacts with perceptions of anonymity to further facilitate perceptions of authenticity or genuine self-expressions. As P3 explained,

> "They [creators] feel like they can be more open because people that can make comments and change what your video is about, they're not seeing it. It's being seen by people who have never met this person. I feel like you could be more of a genuine person that way because your video isn't being, I guess I want to say corrupted, you know what I mean?"

This participant implies that social media content is often shaped by one's assessment of others' expectations, particularly known others' (e.g., significant others, co-workers) expectations. That is, social networks exert a powerful influence on what types of self-expression, opinions, and self-presentations are determined to be acceptable, and potentially in a way that suppresses or overrides the creator's original intention. P15 also noted that strangers' opinions, especially negative reactions, were less impactful than

known others' or friends' opinions. In reflecting on posting images related to personal weight loss, P15 explained:

> "I personally wouldn't post that on my Instagram, just because if my family or friends from high school follow me, I don't really want them to see my before and after pictures. But on TikTok, I didn't really care because I didn't know who was following me. So it just felt more comfortable."

This comment presents an intriguing contradiction for authenticity, in that responses from strangers are not as influential or "corrupting" as responses from known others, which facilitates disinhibited self-expression and ostensibly more authentic self-presentation. That said, and as we explore in the following section, authenticity as a site norm is indeed socially policed, and comments from (unknown) others do shape what content and modes of expression are perceived by creators as likely to be positively received by an audience of strangers. As these experiences suggest, anonymity and association between content afforded by TikTok's features, such as the "For You" page and general emphasis on association between content rather than association between users, are interlinked and potentially inform each other in supporting a space that users perceive to be less judgmental and more accepting of "interesting" and varied content than other social media platforms. Future research could explore these potential connections through additional methods.

4.3.3 Audience and comments as enforcing authenticity. As discussed, social media norms arise from interactions between users as well as between users and material features of a platform. On TikTok, interactions between users commonly occur through heart-shaped "likes", direct messages (private, one-to-one communication), and comments. Unlike platforms such as Instagram and Facebook, comments on TikTok are not immediately visible or previewed on posts, though a link to the comment section is part of every post, by default, in the form of a comment icon (i.e., image of a speech bubble). The number of comments on a TikTok video is visible on that video, as is the number of likes. Users click on the comment icon to access the comment section. This additional step required to view comments provides some separation between both creators and viewers as well as creators/viewers and comments. As P6 summarized, "I feel like because the comments are kind of tucked away in the corner, you kind of have to go out of your way to look at them." This separation was also remarkable in light of the reality that comments on TikTok, as on other social media, can perpetuate hateful judgment, including racist, sexist, homophobic, and transphobic views. As such, a number of participants, such as P1, noted, "in general I just don't check the comments." It should be noted that users may also disable comments for a particular video, providing an additional layer of separation between creator and audience. While none of our participants reported turning off comments, some observed it and perceived it as strategic. As P11 explained,

> "I've noticed that, so that they can't get any negative feedback, because if you are putting yourself out there, I understand why they've turned the comments off. They don't want any negative feedback, which I feel like it makes sense. I would do that too."

Strategic disengagement of comments further highlights the role of comments sections as sites of judgment of and reaction to personal vulnerability as authentic.

While the relative invisibility and likelihood of comments as a site of judgment complicate the meaningfulness of comments on TikTok, some participants did attend to comments in a way that suggests interactions with other users through comments serve to enforce authenticity as a norm. P3, for instance, drew on comments as a guide for content:

> "I think I come across to people as a little bit awkward, but I genuinely want to help people, and people have commented on that. They've looked through my videos and they're just like, 'You have your moments where you're kind of awkward,' which I agree with. I have my moments where I'm trying to show emotion and that you can just tell I'm faking it. …And when it comes to videos where I'm really trying to help somebody, you can see the actual emotion. And they're like, 'We need to see more videos like this because the ones where you're trying to act, your awkwardness comes across.' So, it's like, I think I've put up this persona of...People prefer to see me as my real self, not as faking it or whatever."

In this way, P3 received validation for their "authentic" self directly from their audience via comments. Importantly, P3 aligns awkwardness with faking emotion, such that both they and their audience align awkwardness with faking with inauthenticity. (In other words, this example does not necessarily imply that all awkwardness is inauthentic.) This experience suggests both the reinforcement of emotional rawness as authentic as well as the reinforcement of awkwardness (i.e., faking emotion) as inauthentic; the commenters aid in upholding the norm of authenticity as both raw and effortless, in effect further reifying "just be you" as an attitude guiding self-expression on TikTok.

This functionality extends to enforcing authenticity in a more literal way, that of calling out duplicative content or content "stealers." In interviews, content thieves were generally regarded as more popular accounts that recreated content from smaller, less visible accounts in an effort to maintain or gain attention. As this behavior is viewed as undesirable on TikTok, users may sanction content thieves through comments. P10 explained, "If enough people have seen the smaller creators [who originated the content], they will call them out in the comments when they are creating copies." Cases of duplication are perhaps easier to verify, given a record of visual proof or similar content as a point of comparison.

Determining the authenticity of personal and vulnerable expressions is a more complicated and error-prone process, and, for better or worse, one in which comments are influential. P4 recalled,

> "I've seen people post something, really [personal], 'I lost my brother, he was everything to me. He died. It was the four of us.' And then it was fake. I am like, got to be kidding me. My heart was aching for it. Then you go into comments, it's like, 'This is fake. You should see his other page.' Now, I'm like, what? Why would you do that? I go into comments and make sure it's not fake."

The judgment of other TikTok users, as manifest in comments, may thus be understood as more authentic or grounded in truth than a personal experience related through video. In this case, additional videos provided support for determining such content inauthentic. In cases where such content is not available, other metrics must be used for determining authenticity. P3 recalled an incident in which comments judging authenticity appeared to incorrectly cry "fake":

> "There was a girl that did a video where she was talking about being suicidal, and she is very known for clout chasing. […] And unfortunately, she was actually trying to do an actual video from her heart, expressing the need for help. But because she's known as a clout chaser, a lot of the comments were just, 'This is fake, this is clickbait, blah, blah, blah. She's not like that.'"

In this way, one's persona, cultivated through content, can interfere with the ability to "authentically" communicate emotional distress, and by extension social support needs, particularly when such distress seems out of character. This friction underscores the difficulty individuals may face in appearing as complex beings on social media; consistency in presentation becomes a double-edged sword in which one

comes to be "read" as authentic through congruent self-presentations, yet deviations from one's typical self-presentation are read as inauthentic, despite conformity with one's authentic experience or true self. Thus, normative authenticity on TikTok may require or be enhanced by consistency in self-expression. Indeed, Gilpin et al. [27] note that "narrative coherence" is required for shifting identities to be determined authentic. More broadly, consistency of communication may be influential in inferring group norms [61] and, when observed, is evidence of group norms functioning [62]; inconsistent communication may inadvertently challenge established group norms or raise uncertainty in others in terms of how they should behave in a group. This example thus underscores the social aspect of authenticity, as "clout chaser" may be an authentic identity (and reflect an authentic motivation for using social media) of an individual, but one that is perceived as undesirable and inauthentic by one's audience. We explore how this example complicates authenticity's implications as a norm in the Discussion.

Participants also suggested that engaging with comments on their videos functioned to support authenticity and to perpetuate their own values and hopes about the platform. P12, for example, commented that not getting responses to comments could be a disappointing experience. "So I try to respond just to show that I care," they said. "I'm an artist and yeah, that doesn't make me anything but human and I care and bleed just like you." P5 also valued showing care, saying, "I want them [my followers] to know that I care and that I see what they comment on my videos." Acting as they hoped others might, by interacting with commenters, appeared to cement perceptions of self (e.g., I am a caring human) as well as reinforce authenticity through self-presentation (e.g., showing others that I am a caring human).

Responding to comments from the audience further underscored authenticity by encouraging ongoing engagement and interaction between the posters and their audience. "I feel like if you show that you appreciate it, they'll continue to engage with you," P15 said. Similarly, P3 explained,

> "I feel like if they're going to comment and they're going to like then they found something important in the video. Then, you know, if they found something that they need to speak on or something like that, then I should engage with that because that's the whole point of the video."

Across these responses is a sense that commenting in a genuine way is an effortful and deliberate act. In this sense, authenticity perceived by viewers and commenters encourages authenticity of both content and in interactions between content creators and viewers, and perpetuates authenticity as a norm governing multiple behaviors (e.g., content creation, interaction) on TikTok.

4.3.4 Comments as a site of social support. In response to RQ3, we find that the mutual influence and reinforcement of authenticity of expression undergirds much of the potential for TikTok to be a site of social support exchange. Genuine or authentically appreciative and supportive comments facilitate connections between users with similar interests and experiences, which "makes you feel like you're not so lonely, and there's so many people out there who relate to you as well," as P7 explained. Finding similar others is a powerful basis for social support [73], [85]. For some, the ability to find relatable content, as afforded by association between content and the "For You" page algorithm, as previously discussed, intersected with perceptions of anonymity (both in terms of discursive anonymity and an audience of strangers) to reduce inhibitions and fear of judgment; this appeared to also result in supportive exchanges. As P11 explained, "Strangers can be so positive and uplifting, and sometimes they'll have similar life experiences...It's nice because they know nothing about your life. You can talk about this one specific thing and that's it. They can't judge you." This example helps to illustrate the connections among anonymity (both discursive and of one's audience), proximate experience, social support, and perceived risk of

expression; in combination, these quotes suggest that comments simply relating to a personal experience may be effective in challenging feelings of isolation and providing meaningful validation.

Indeed, in discussions of TikToks on sensitive personal experiences (e.g., addiction, abuse, sexual assault), participants reflected that the majority of comments appeared to be supportive. As P15 reflected, "Usually, when someone finally makes a more personal content, all the comments are supportive." They continued, "I've never seen a negative comment on a personal video like that [about sexual assault]." This comment contrasts with the experience of P3, previously related, regarding the treatment of a "clout chaser" expressing difficult emotions. A comment from P8 intimates that motivation of expression may also inform assessments of authenticity:

> "I don't think many people would be vulnerable in a video and cry if they just wanted likes, if they just wanted followers. I mean, that's something you're sharing pretty raw emotions. Most people are going to do that and they're genuine."

Thus, expressions of raw and difficult emotions may be read as authentic if interpreted to be motivated by a desire to vent, seek support, or raise awareness, while expressions motivated by an interest in growing one's audience may be read as inauthentic. While a detailed analysis of motivation as a criterion for evaluating authenticity of expression is beyond the scope of this paper, we note it here to further outline audience's role in enforcing and sanctioning authenticity as well as to highlight evaluation of authentic expression on social media as an area for future work.

In interpreting responses perceived as offering support, participants also alluded to authenticity in responses themselves. P12, for example, explained,

> "I see a lot of those people [in comments] relating to the topic at hand. And I also see a lot of people trying to be inspiring. Trying to uplift the person. Like I said, the 'you're not alone in this. We're all here for you. We care about you. We love you.' [...] Because they don't like seeing a person in pain."

In this case, supportive comments are attributed to a genuine, empathic response to witnessing distress, and thus may authentically reflect a desire to support the creator. (Whether sentiments such as love for an internet stranger are similarly authentic emotional expressions is a slightly different question.) P11 spoke to another way that authenticity influenced such interactions, saying,

> "I feel like it's important to comment because these people are putting out details of their life and the struggles that they went through. They're sacrificing some of their own privacy and going out of their comfort zone to post these things. I feel like these people should be appreciated because things like this happen in people's lives and the fact that they're sharing it for other people to see means a lot to me."

In this comment, P11 acknowledges the vulnerability associated with sharing difficult emotions and experiences and intimates a responsibility to assume such content is authentic and engage with it accordingly. Similarly, P8 expressed a value in sharing personal stories and finding similar others that also assumes such stories are authentic:

> "I think that's how we develop empathy and compassion is, we learn people's stories, and we learn that they have things in common with us, and they have struggles that are horrible, and we can identify with them."

In these examples, authentic self-presentation and finding commonalities, including difficult emotions, on TikTok facilitates empathic communication; this connection underscores the potential relationship between normative authenticity and social support.

Although the examples included here contrast with instances of harassment informed by commenters' determinations of in/authentic content, we view the potential for and demonstrated presence of social support on TikTok as encouraging. As P14 summarized, "There's always going to be people who are trolling, A-holes that really think they need to ruin everyone's day. But there's a lot of really pure and good heartfelt comments there too." Indeed, these examples make apparent the viability of strangers as support providers, as well as underscore the role of interactivity (e.g., through comments) in constructing and maintaining normative authenticity, both in policing seemingly inauthentic content and in affirming emotionally raw content as authentic through provision of authentically (i.e., genuinely) supportive comments.

5 DISCUSSION

We make the following contributions to the CSCW and social computing literature:

- An in-depth understanding of how the self-presentation norm of authenticity is learned, enacted, and enforced on TikTok
- Identify the sociotechnical affordances that support authenticity as a self-presentation norm on TikTok
- Theorize connections between the construction of authenticity, sharing both positive and difficult emotions, and social support, and providing implications for designing social media to facilitate social support exchange

We argue that the social and material factors of TikTok support authenticity as a self-presentational norm. As a norm, authenticity may result in disinhibited self-presentation that feels authentic to users as well as provide a lens through which to filter self-presentation; we reiterate that "authentic" need not be synonymous with "unfiltered," but note that this is one way authenticity may manifest on TikTok. We also argue that the factors that contribute to normative authenticity also contribute to acceptance of a broader range of emotional expression, particularly difficult emotions, than may be perceived on other social media, such as Instagram. More specifically, we argue that user perceptions and TikTok's own branding of the platform as a space for creativity and "fun" support an overarching attitude of "just be you" that also serves to establish authenticity as a self-presentation norm. TikTok's algorithm and design afford association between content, rather than users, as well as perceptions of anonymity of both oneself and one's audience; these affordances, we argue, facilitate finding similar others (e.g., similar identities or experiences) and may encourage disinhibited communication due to reduction of perceived risk associated with self-expression. These outcomes indicate the potential for TikTok as a site for social support requests and provision, and our findings regarding the comments section as a site of both norm judgment/enforcement and support provision provide further evidence of this potential.

The ability of a particular configuration of social media features, affordances, and norms to support a user's perception that they can be themselves, share difficult emotion, and share interesting or mundane or intimate personal experiences—without a constraining fear of judgment from others—carries remarkable implications for designing technology for social support. In the remainder of this section, we argue that the intersection of three factors—perceived anonymity, association between content, and video

modality of TikTok—is an especially fruitful one to consider in this regard. We then consider the limits of normative authenticity and for whom such authenticity is available. Limitations of the present study follow.

5.1 Material factors facilitating authenticity and social support

How norms arise from sociomaterial factors of social media platforms is critical to examine regarding authenticity on social media as well as designing spaces intended or used for social support exchange. This study contributes to both bodies of work by 1) identifying user perceptions, platform features, and affordances that work together to create and uphold normative authenticity on TikTok; and 2) mapping the aforementioned factors to social support processes. That norms vary across social media platforms is well established in research literature [57], [75], [88]; we leverage Waterloo et al.'s [80] understanding of social context as composed of social and material factors and an affordances lens to identify specific factors—anonymity, association, video modality—and attitudes—"just be you," "fun" content—that contribute to authenticity as social norm. Our findings demonstrate not only the interactions between social and material factors, but also the ways in which affordances may interact with each other. Building on our work, future research could explore these interactions in more depth to present an affordance-based model of norm formation on social media.

Anonymity, both of content creators and of one's audience, potentially affects the intimacy and depth of personal expression and disclosure of personal experiences on TikTok. That is, we argue that anonymity upholds normative authenticity on TikTok by reducing the risks associated with personal expression and disclosure. In online contexts, anonymity has been associated with a disinhibition effect, in which individuals may express themselves more freely and perhaps with more intimacy of depth of disclosure, than in a face-to-face, offline context, if they believe their identity to be anonymous [72]. Furthermore, believing oneself to be anonymous to an audience, or further, as in this case, to an audience of strangers, may lower perceived risks, such as negative judgment, associated with disclosure [5], [72]. As one participant explained, negative consequences arising from TikToks are generally limited to "bad comments." We address this generalization further momentarily. Indeed, both discursive anonymity of oneself and anonymity of one's audience appears to affect risk perception, as strangers who "don't know you" are deemed unable to "judge" you. Consequently, both anonymity of self and anonymity of audience may contribute to the viability of "just be you" as a behavioral guideline on TikTok, and thus uphold authenticity as a self-presentation norm.

We found that association between content, rather than between users, further supports perceived anonymity and upholds "just be you" as an attitude on TikTok. In combination, anonymity (of self and of others) and association between content may facilitate reduced risks associated with expression, finding experientially similar or like-minded others, and potentially encourage empathy in responding to individuals sharing difficult emotions. In this way, association, afforded by network structure and the "For You" page algorithm, and anonymity are mutually influential in affecting perceived risks associated with content sharing, including emotional expression, and by extension, mutually influential in supporting authenticity as a self-presentation norm. We also note that association between content allows users to connect with content on the basis of similar interests or shared experiences, which can provide a basis for social support and empathy [73], [85]. A caveat to this, however, is the notion that anonymity may undermine credibility in support contexts, in that individuals may worry about or be susceptible to taking

bad advice or taking a disingenuous suggestion seriously [9]. This risk may be assuaged by a third factor, modality of content.

The video modality of TikTok also interacts with and affects site norms, in that it is common for users to be present in their videos; many TikToks simply consist of a creator facing the camera and speaking directly to an audience. As noted, video may provide "proof" of emotion or experience [32], [70], which may in turn be influential in assessing authenticity. The video modality of TikTok may support a higher degree of nonverbal cues than other modes of content. This richness potentially supports authentic emotional expression and evaluation thereof by providing dynamic visual evidence of emotional states (e.g., crying), experiences (e.g., scars, hospital), and oneself. As our findings show, this visual evidence may encourage interpretation of such content as authentic as well as facilitate empathic responses and social support provided via comments.

Indeed, video presence that simulates eye contact may also reduce negative disinhibited communication, such as flaming or harassment [48], which may also encourage empathic response to emotional expression. (As participants indicated that comments on TikTok still perpetuated harassment, however, the effects of simulated eye contact via asynchronous video communication on disinhibited communication warrants further study.) By extension, we argue that TikTok may be especially well-suited for *providing* support to viewers via nonverbal immediacy cues in videos. This may be particularly impactful for "lurkers," or those who do not directly engage with content via likes, comments, or messages, as viewing a high-immediacy message may not require direct solicitation of support. These connections warrant further study, as emerging work already indicates the potential of TikTok in effectively communicating other forms of health messages and resources [90].

We conclude that social media that relies on hyper-visual content, such as video; that affords anonymity through features, like pseudonymous usernames, and affordances, like privileging association between content over association between users (which further supports anonymity, within networks); *and* that affords association based on shared experience or similar interest, could facilitate computer-mediated social support in a way that other platforms have less successfully captured. On TikTok, these factors interact with each other as well as with other perceptions of the platform (e.g., youth-oriented, mundane topics) to support authenticity as a self-presentational norm, which in turn appears to challenge (or at least not recreate) emotional sharing norms such as social positivity bias and instead value expression of both positive and difficult emotions.

This study also contributes to the argument that social factors are also design considerations. It supports the idea that social norms can be influential in bounding spaces conducive to disclosure and support seeking. Further research could explore social support exchange on TikTok specifically through an affordance lens, including how exchange might occur outside of the comments field. Further research is also needed to explore the efficacy of masspersonal video messages as vehicles for supportive communication. Additionally, interaction via comments and messages continue to be vexing from a social support design perspective, given the potential of comments as a means of providing emotional and informational support as well as harassment and vitriol. Constraining the visibility of comments, as discussed in our findings, may help in providing a barrier between creators and comments, such that creators may limit their exposure to potentially harmful comments. We suggest that future scholarship on designing digital spaces for social support emphasize the importance of interaction for social support as well as authentic self-presentation, and design mechanisms for interaction with these considerations in mind.

5.2 Authenticity and identity

Though not an emphasis of this study, for whom authenticity is available and how authenticity is evaluated remain crucial considerations. Our small sample size limits the generalizability of our findings and restricts our ability to discern if differences in perceptions of authenticity are individual in nature or indicative of broader differences that recreate social biases and marginalization. In other words, what is considered authentic expression, how it is evaluated, and by whom it is evaluated likely differ across communities and social identities, as well as across platforms. Indeed, Haimson and Hoffmann [33] argue that authenticity, as it pertains to one's identity on Facebook, is not viable for users with shifting or non-normative identities, such as transgender people and survivors of abuse. Others have argued that while individuals marginalized along axes of race, gender, sexuality, ability, and/or body size [43] or gender and sexuality [69] are able to form solidarity on TikTok, they also face what is referred to as "algorithmic symbolic annihilation" [43] (i.e., how algorithms perpetuate normative narratives about phenomena in which what is accounted for has power, and what is not does not [4]) or similar concepts such as "algorithmic exclusion" [69]. Like all design choices, the policies and features that shape identity expression on social media are political [83]. While our findings do not speak directly to experiences of marginality, the experience related by P3 regarding a "clout chaser" indicates that social evaluation of authenticity (i.e., by other users) is fallible and may rely on imperfect evaluative criteria, such as consistency of expression.

Despite the attitudes expressed by some participants that a negative comment is a minor consequence of content creation, harassment via social media constitutes real harm, often against members of marginalized communities, and this harassment may be facilitated by factors such as anonymity and disinhibited communication [31]. Beyond comment sections, there is a "dark side" to TikTok [44], [81] that also warrants further scholarly consideration. The sociomaterial factors that afford beneficial freedom of emotional expression may similarly afford expression of "authentic" (i.e., reflective of one's beliefs) ideological views (e.g., homophobia, transphobia, sexism, racism, white supremacist attitudes) that pose harm. Relatedly, the association afforded by features such as the "For You" page may contribute to the formation of new social networks that share these views and unintentionally amplify them to the detriment of other users and social groups. In designing spaces for connection and social support, potential consequences such as facilitating authentic expression of hostile ideologies must be taken into account and is an area for future work. This is not to say that designing with authenticity in mind is without merit, but to instead highlight the additional components of platform design and context, such as contend moderation and policy development, that are influential in this regard.

5.3 Limitations

In considering how the norm of authenticity is supported by TikTok's features, affordances, and users, this paper takes a narrow view of the platform. There are certainly additional aspects of TikTok that differentiate it from other platforms, such as a young user base, that further shape social norms, expectations, and perceptions of the platform. In other words, the factors we identified may not fully account for the formation and perpetuation of authenticity as a norm of self-expression, and future research could explore the intersections among anonymity, association, and content modality in more depth. We also did not explore a prominent expression norm of TikTok, the mimetic use of "sounds" by multiple users and lip-synching, and how the replicability afforded by this feature affects perceptions of authentic self-expression by the creator and audience. Similarly, content creation mechanisms on TikTok

afford editability of videos, such that creators can add effects, transitions, and otherwise alter raw footage for posting. Beyond content modality, modes of storytelling likely also impact expectations and perceptions of authenticity on social media and warrant further scholarly consideration. While generalizability is not the goal of in-depth interview studies, an additional limitation of this study is the small number of interview participants, which limits the generalizability of findings. However, participant responses suggest the potential of TikTok for supportive communication; future research could explore this potential by focusing more explicitly on the myriad communities intentionally engaged in destigmatizing, normalizing, and providing mutual support for difficult experiences on TikTok. Additionally, we did not differentiate between practices of content creation and consumption; future research could explore the distinctions between user experiences of creating TikTok content and consuming content, and how such practices affect perceptions of norms and social support on the platform. Finally, while experiences of harassment did not surface prominently in our data, we acknowledge that comments on TikTok may also facilitate harm via harassment and interpersonal judgment that may be further impacted (amplified or ameliorated) by the video modality of TikTok and visual identification of content creators. We did not screen participants based on positive or negative experiences on TikTok, but suggest that examining experiences of harassment on TikTok and the role of the affordances we identified in this study is an area for future work.

6 CONCLUSION

We draw on users' experiences with TikTok broadly to consider how authenticity, as a self-presentational norm, is learned, constructed, and enacted on TikTok. Drawing on the factors that comprise social context as well as an affordance lens, we identify material factors—the "For You" page, policies allowing pseudonymous usernames, and video modality—and sociotechnical affordances—perceived anonymity (of oneself and one's audience) and association between content—that support normative authenticity on TikTok. We find that these affordances, in combination with a "just be you" attitude, inform user perception of both goofy content and "raw" emotionality as authentic. This range of acceptable emotionality (i.e., from goofy to difficult) suggests that normative authenticity on TikTok may make the platform conducive to both the expression of difficult emotions and experiences leading to social support exchange. Our findings provide preliminary evidence of user comments as a site of norm judgment and sanctioning as well as social support provision. In identifying sociomaterial factors contributing to authenticity on TikTok, this paper illustrates how an affordance lens may be used to trace norm development and perpetuation on social media. We identify avenues for future work, including analysis of how identity and marginality affect perception, judgment, and sanctioning of normative authenticity. We discuss implications of normative authenticity for designing social media for social support.

ACKNOWLEDGMENTS

We thank the participants of this study for sharing their thoughts and experiences with us. We also thank Nadia Karizat, MSc for her contributions to the study and data collection that informs this paper.

REFERENCES

[1] Nazanin Andalibi. 2020. Disclosure, Privacy, and stigma on social media: Examining non-disclosure of distressing experiences. *ACM Trans. Comput.-Hum. Interact.* 27, 3, Article 18 (June 2020), 43 pages. DOI: https://doi.org/10.1145/3386600

[2] Nazanin Andalibi and Andrea Forte. 2018. Announcing pregnancy loss on Facebook: A decision-making framework for

stigmatized disclosures on identified social network sites. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18),* April 21–26, 2018, Montreal QC, Canada. ACM, New York, NY, Paper 158, 1–14. DOI: https://doi.org/10.1145/3173574.3173732

[3] Nazanin Andalibi and Andrea Forte. 2018. Responding to sensitive disclosures on social media. *ACM Trans. Comput.-Hum. Interact.* 25, 6, Article 31 (December 2018), 29 pages. DOI: https://doi.org/10.1145/3241044

[4] Nazanin Andalibi and Patricia Garcia. 2021. Sensemaking and coping after pregnancy loss: The seeking and disruption of emotional validation online. In *PACM on Human Computer Interaction*, 5, CSCW1, Article 127, April 2021. New York, NY. 31 pages. DOI: https://doi.org/10.1145/3449201

[5] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2018. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Trans. Comput.-Hum. Interact.,* 25, 5, Article 28 (October 2018), 35 pages. DOI: https://doi.org/10.1145/3234942

[6] Anonymous. (1998). To reveal or not to reveal: A theoretical model of anonymous communication. *Communication Theory,* 8, 4, 369 475 DOI https //doi org/10 1111/j 1468 2885 1998 tb00226 x

[7] Kristine Ask and Crystal Abidin. 2018. My life is a mess: Self-deprecating relatability and collective identities in the memification of student issues. *Information, Communication & Society,* 21, 6, 834–850. DOI: https://doi.org/10.1080.1369118X.2018.1437204.

[8] Godfrey T. Barrett-Lennard. 1998. *Carl Rogers' Helping System: Journey and Substance.* Sage, London, UK.

[9] Kristen Barta. 2019. Reclaiming Publicness in the Face of Sexual Assault: Social Media, Disclosure, and Visibility (Communication). PhD Dissertation. University of Washington, Seattle, WA.

[10] Joseph B. Bayer, Penny Trieu, and Nicole B. Ellison. 2020. Social media elements, ecologies, and effects. *Annual Review of Psychology,* 71, 471–497. DOI: https://doi.org/10.1146/annurev-psych-010419-050944

[11] Natalya N. Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '15),* March 14–18, 2015, Vancouver BC, Canada. ACM, New York, NY, 154–164. DOI: https://doi.org/10.1145/2675133.2675297

[12] Elena Botella. 4 Dec 2019. TikTok admits it suppressed videos by disabled, queer, and fat creators. *Slate.* Retrieved from https://slate.com/technology/2019/12/tiktok-disabled-users-videos-suppressed.html

[13] Moira Burke, Cameron Mariow, and Thomas Lento. 2009. Feed me: Motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09).* ACM, New York, NY, 945–954. DOI: https://doi.org/10.1145/1518701.1518847

[14] Brant. R. Burleson and Erina L. MacGeorge. 2002. Supportive communication. In *Handbook of Interpersonal Communication*, 3rd ed., Mark L. Knapp and John A. Daly, Eds. Sage, Thousand Oaks, CA, 374–424.

[15] Caleb T. Carr and Rebecca A. Hayes. 2015. Social media: Defining, developing, and divining. *Atlantic Journal of Communication,* 23, 1, 46–65. DOI: https://doi.org/10.1080/15456870.2015.972282

[16] Caleb T. Carr, D. Yvette Wohn, and Rebecca A. Hayes. 2016. Like as social support: Relational closeness, automaticity, and interpreting social support from paralinguistic digital affordances in social media. *Computers in Human Behavior,* 62, 385–393. DOI: https://doi.org/10.1016/j.chb.2016.03.087

[17] Robert B. Cialdini and Melanie R. Trost. 1998. Social influence: Social norms, conformity and compliance. In *The Handbook of Social Psychology*, Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (Eds.). McGraw Hill, Boston, MA, 151–192.

[18] Jenny L. Davis and James B. Chouinard. 2016. Theorizing affordances: From request to refuse. *Bulletin of Science, Technology & Society,* 36, 4, 241–248. DOI: https://doi.org/10.1177/0270467617714944

[19] Sofia Dewar, Schinria Islam, Elizabeth Resor, and Niloufar Salehi. 2019. Finsta: Creating "fake" spaces for authentic performance. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19).* ACM, New York, NY, Paper LBW1214, 1–6. DOI: https://doi.org/10.1145/3290607.3313033

[20] Nicole B. Ellison, Lindsay Blackwell, Cliff Lampe, and Penny Trieu. 2016. "'The question exists, but you don't exist with it": Strategic anonymity in the social lives of adolescents. *Social Media and Society,* 2, 4. DOI: https://doi.org/10.1177/2056305116670673

[21] Nicole B. Ellison and danah m. boyd. (2013). Sociality through social network sites. In *The Oxford Handbook of Internet Studies.* Oxford University Press, Oxford, UK, 151–172.

[22] Sandra K. Evans, Katy E. Pearce, Jessica Vitak, and Jeffrey W. Treem. 2017. Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication,* 22, 1, 35–52. DOI: https://doi.org/10.1111/jcc4.12180

[23] Leon Festinger (1954) A theory of social comparison processes *Human Relations,* 7, 2, 117 140 DOI https://doi.org/10.1177/001872675400700202

[24] Amanda L. Forest and Joanne V. Wood. 2012. When social networking is not working: Individuals with low self-esteem

recognize but do not reap the benefits of self-disclosure on Facebook. *Psychological Science*, 23, 3, 295–302. DOI: https://doi.org/10.1177/0956797611429709

[25] Jesse Fox and Bree McEwan. 2017. Distinguishing technologies for social interaction: The perceived social affordances of communication channels scale. *Communication Monographs*, 84, 3, 298–318. DOI: https://doi.org/10.1080/03637751.2017.1332418

[26] Georgia Gaden and Delia Dumitrica. 2015. The "real deal": Strategic authenticity, politics and social media. *First Monday*, 20, 1 (January 2015). DOI: https://doi.org/10.5210/fm.v20i1.4985

[27] Dawn R. Gilpin, Edward T. Palazzolo, and Nicholas Brody. 2010. Socially mediated authenticity. *Journal of Communication Management*, 14, 3, 258–278. DOI: https://doi.org/10.1108/13632541011064526

[28] Barney G. Glaser and Ansel L. Strauss. 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine, Chicago, IL.

[29] Daena J. Goldsmith. 2004. *Communicating Social Support*. Cambridge University Press, Cambridge, UK.

[30] Mark S. Granovetter. 1973. The strength of weak ties. *American Journal of Sociology*, 78, 6, 1360–1380. Retrieved from https://www.jstor.org/stable/2776392

[31] Kishonna L. Gray. 2012. Intersecting oppressions and online communities: Examining the experiences of women of color in Xbox Live. *Information, Communication and Society*, 15, 3, 411–428. DOI: https://doi.org/10.1080/1369118X.2011.642401

[32] Gabrielle Grow and Janelle Ward. 2013. The role of authenticity in electoral social media campaigns. *First Monday*, 18, 4 (April 2013) DOI https //doi org/10 5210/fm v18i4 4269

[33] Oliver L. Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday*, 21, 6 (June 2016). DOI: https://doi.org/10.5210/fm.v21i6.6791

[34] Hanno Hardt. 1993. Authenticity, communication, and critical theory. *Critical Studies in Mass Communication*, 10, 1, 49–69. DOI: https://doi.org/10.1080/15295039309366848

[35] Rebecca A. Hayes, Caleb T. Carr, and Donghee Yvette Wohn. 2016. One click, many meanings: Interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting and Electronic Media*, 60, 1, 171–187. DOI: https://doi.org/10.1080/08838151.2015.1127248

[36] Natalie Ann Hendry. 2020. Young women's mental illness and (in-)visible social media practices of control and emotional recognition. *Social Media + Society*, 6, 4. DOI: https://doi.org/10.1177/2056305120963832

[37] Eren E. Hollenbaugh and Marcia K. Everett. 2013. The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, 18, 3, 283–302. DOI: https //doi org/10 1111/jcc4 12008

[38] Martin Holt and Christine Griffin. 2003. Being gay, being straight and being yourself. *European Journal of Cultural Studies*, 6, 3, 404–425. DOI: https://doi.org/10.1177/13675494030063008

[39] How TikTok recommends videos #ForYou. 18 June 2020. TikTok. Retrieved from https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

[40] Harry M. Johnson. 1960. *Sociology: A Systematic Introduction*. Harcourt, Brace and World, New York, NY.

[41] Katrina P. Jongman-Sereno and Mark R. Leary. 2019. The enigma of being yourself: A critical examination of the concept of authenticity. *Review of General Psychology*, 23, 1, 133–142. DOI: https://doi.org/10.1037/gpr0000157

[42] Akane Kanai. 2019. On not taking the self seriously: Resilience, relatability and humour in young women's Tumblr blogs. *European Journal of Cultural Studies*, 22, 1, 60–77. DOI: https://doi.org/10.1177/1367549417722092.

[43] Nadia Karizat, Dan Delmonaco, Motahare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. In *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '21)*. October 23–27, held virtually. ACM, New York, NY, 26 pages.

[44] Melanie Kennedy. 2020. "If the rise of the TikTok dance and e-girl aesthetic has taught us anything, it's that teenage girls rule the internet right now": TikTok celebrity, girls and the Coronavirus crisis. *European Journal of Cultural Studies*, 23, 6, 1069–1076. DOI: https://doi.org/10.1177/1367549420945341

[45] Michael. H. Kernis and Brian. M. Goldman. 2006. A multicomponent conceptualization of authenticity: Theory and research. *Advances in Experimental Social Psychology*, 38, 6, 283–357. DOI: https://doi.org/10.1016.S0065-2601(06)38006-9

[46] Robert E. Kraut and Paul Resnick. 2011. Building Successful Online Communities: Evidence-based Social Design. MIT Press, Cambridge, MA.

[47] Venla Kuuluvainen and Pekka Isotalus. 2015. Words and beyond: Members' experiences of the supportive communication and helping mechanisms of Al-Anon groups. *Journal of Groups in Addiction & Recovery*, 10, 3, 204–223. DOI: https://doi.org/10.1080/1556035X.2015.1066725

[48] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online

disinhibition. *Computers in Human Behavior,* 28, 2, 434–443. DOI: https://doi.org/10.1016/j.chb.2011.10.014

[49] Mark R. Leary. 1996. Self-presentation: Impression Management and Interpersonal Behavior. Westview Press, Boulder, CO.

[50] Mark R. Leary and June Price Tangney. 2003. The self as an organizing construct in the behavioral and social sciences. In *Handbook of Self and Identity,* Mark R. Leary and June Price Tangney (Eds.). Guilford Press, New York, NY, 3–14.

[51] Thomas R. Lindlof and Bryan C. Taylor. 2011. *Qualitative Communication Research Methods* (3rd ed). Sage, Los Angeles, CA.

[52] Eden Litt. 2012. Knock, knock. Who's there? The imagined audience. *Journal of Broadcasting and Electronic Media,* 56, 3, 330–345. DOI: https://doi.org/10.1080/08838151.2012.705195

[53] Bingjie Liu and Jin Kang. 2017. Publicness and directedness: Effects of social media affordances on attributions and social perceptions. *Computers in Human Behavior,* 75, 70–80. DOI: https://doi.org/10.1016/j.chb.2017.04.053

[54] Xiao Ma, Nazanin Andalibi, Louise Barkhuus, and Mor Naaman. 2017. "People are either too fake or too real": Opportunities and challenges in tie based anonymity In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17*), May 6–11, 2017, Denver, CO. ACM, New York, NY, 1781–1793. DOI: https://doi.org/10.1145/3025453.3025956

[55] Alice E. Marwick and danah boyd. 2010. "I tweet honestly, I tweet passionately": Twitter users, context collapse, and the imagined audience. *New Media and Society,* 13, 1, 114–133. DOI: https://doi.org/10.1177/1461444810365313

[56] Megan McCluskey. 22 July 2020. Black TikTok creators say their content is being suppressed. *Time.* Retrieved from https://time.com/5863350/tiktok-black-creators/

[57] Caitlin McLaughlin and Jessica Vitak. 2012. Norm evolution and violation on Facebook. *New Media and Society,* 14, 2, 299–315. DOI: https://doi.org/10.1177/1461444811412712

[58] Kembrew McLeod. 1999. Authenticity within hip-hop and other cultures threatened with assimilation. *Journal of Communication,* 49, 4, 134–150. DOI: https://doi.org/10.1111/j.1460-2466.1999.tb02821.x

[59] Juan-Carlos Molleda. 2010. Authenticity and the construct's dimensions in public relations and communication research. *Journal of Communication Management,* 14, 3, 223–236. DOI: https://doi.org/10/1108/13632541011064508

[60] Our mission. 2020. *TikTok.* Retrieved from https://www.tiktok.com/about?lang=en

[61] William Foster Owen. 1985. Metaphor analysis of cohesiveness in small discussion groups. *Small Group Research,* 16, 3, 415–424. DOI: https://doi.org/10.1177/0090552685163011

[62] Tom Postmes, Russell Spears, and Martin Lea. 2000. The formation of group norms in computer-mediated communication. *Human Communication Research,* 26, 3, 341–371. DOI: https://doi.org/10.1111/j.1468-2958.2000.tb00761.x

[63] Hua Qian and Craig R. Scott. 2007. Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication,* 12, 4, 1428–1451. DOI: https://doi.org/10.1111/j.1083-6101.2007.00380.x

[64] Leonard Reinecke and Sabine Trepte. 2014. Authenticity and well-being on social network sites: A two-wave longitudinal study on the effects of online authenticity and the positivity bias in SNS communication. *Computers in Human Behavior,* 30, 95–102. DOI: https://doi.org/10.1016/j.chb.2013.07.030

[65] Greg Roumeliotis, Yingzhi Yang, Echo Wang, and Alexandra Alper. 1 Nov. 2019. Exclusive: U.S. opens national security investigation into TikTok - sources. *Reuters.* Retrieved from https://www.reuters.com/article/us-tiktok-cfius-exclusive/exclusive-u-s-opens-national-security-investigation-into-tiktok-sources-idUSKBN1XB4IL

[66] Lauren Scissors, Moira Burke, and Steven Wengrovitz. 2016. What's in a like?: Attitudes and behaviors around receiving likes on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16*), ACM, New York, NY, 1501–1510. DOI: https://doi.org/10.1145/2818048.2820066

[67] Susie Scott. (2004). Researching shyness: A contradiction in terms? *Qualitative Research,* 4, 1, 91–105. DOI: https://doi.org/10.1177/1468794104041109

[68] Kennon M. Sheldon, Richard M. Ryan, Laird J. Rawsthorne, and Barbara C. Ilardi. 1997. Trait self and true self: Cross-role variation in the big-five personality traits and its relations with psychological authenticity and subjective well-being. *Journal of Personality and Social Psychology,* 73, 6, 1380–1393. DOI: https://doi.org/10/1037/0022-3514.73.6.1380

[69] Ellen Simpson and Bryan Semaan. 2020. For you, or for "you"?: Everyday LGBTQ+ encounters with TikTok. *Proc. ACM Hum.-Comput. Interac.,* 4, CSCW3, Article 252 (December 2020), 34 pages. DOI: https://doi.org/10.1145/3432951

[70] Susan Sontag. 1977. *On Photography.* Farrar, Straus and Giroux, United Kingdom.

[71] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology—an overview. In *Handbook of Qualitative Research,* Norman K. Denzin and Yvonna S. Lincoln (Eds). Sage, Thousand Oaks, CA, 273–285.

[72] John Suler. 2004. The online disinhibition effect. *Cyberpsychology and Behavior,* 7, 3, 321–326. DOI: https //doi org/10 1089/1094931041291295

[73] Peggy A. Thoits. 2011. Mechanisms linking social ties and support to physical and mental health. *Journal of Health and*

Social Behavior, 52, 2, 145–161. DOI: https://doi.org/10.1177/0022146510395592

[74] Jeffrey W. Treem, and Paul M. Leonardi. 2012. Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Communication Yearbook*, 36, 143–189. DOI: https://doi.org/10.2139/ssrn.2129853

[75] Suvi Uski and Airi Lampinen. 2016. Social norms and self-presentation on social network sites: Profile work in action. *New Media and Society*, 18, 3, 447–464. DOI: https://doi.org/10.1177/1461444814543164

[76] Sonya Utz. 2011. Social network site use among Dutch students: Effects of time and platform. In *Networked Sociability and Individualism: Technology for Personal and Professional Relationships*, Francesca Comunello (Ed.), IGI Global, 103–125. DOI: https://doi.org/10.4018/978-1-61350-338-6.ch006

[77] Jacqueline Ryan Vickery. 2015. "I don't have anything to hide, but … ": The challenges and negotiations of social and mobile media privacy for non-dominant youth. *Information Communication and Society*, 18, 3, 281–294. DOI: https://doi.org/10.1080/1369118X.2014.989251

[78] Jessica Vitak and Jinyoung Kim. 2014. "'You can't block people offline." In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*, February 15–19, 2014, Baltimore, MD, USA. ACM, New York, NY, 461–474. DOI: https://doi.org/10.1145/2531602.2531672

[79] Erin A. Vogel, Jason P. Rose, Lindsay R. Roberts, and Katheryn Eckles. 2014. Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture*, 3, 4, 206–222. DOI: https://doi.org/10.1037/ppm0000047

[80] Sophie F. Waterloo, Susanne E. Baumgartner, Jochen Peter, and Patti M. Valkenburg. 2018. Norms of online expressions of emotion Comparing Facebook, Twitter, Instagram, and WhatsApp *New Media and Society*, 20, 5, 1813 1831 DOI https://doi.org/10.1177/146144481770734

[81] Gabriel Weimann and Natalie Masri. 2020. Research note: Spreading hate on TikTok. *Studies in Conflict and Terrorism*, 14 pages. DOI: https://doi.org/10.1080/1057610X.2020.1780027

[82] Timothy D. Wilson and Elizabeth W. Dunn. 2004. Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, 55, 1, 493–518. DOI: https://doi.org/10.1146/annurev.psych.55.090902.141954

[83] Langdon Winner. 1986. Do artifacts have politics? In *In the Whale and the Reactor: A Search for Limits in an Age of High Technology*, Langdon Winner (Ed.) University of Chicago Press, Chicago, IL, 19–39.

[84] Alex M. Wood, Alex P. Linley, John Maltby, Michael Baliousis, and Stephen Joseph. 2008. The authentic personality: A theoretical and empirical conceptualization and the development of the authenticity scale. *Journal of Counseling Psychology*, 55, 3, 385–399. DOI: https://doi.org/10.1037/0022-0167.55.3.385

[85] Kevin B. Wright, Amy Janan Johnson, Daniel R. Bernard, and Joshua Averbeck. 2011. Computer-mediated social support: Promises and pitfalls for individuals coping with health concerns In *The Routledge Handbook of Health Communication* (2nd ed.), Teresa L. Thompson, Roxanne Parrott, and Jon F. Nussbaum (Eds.). Routledge, New York, NY, 349–362.

[86] Sijia Xiao, Danae Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. 2020. Random, messy, funny, raw: Finstas as intimate reconfigurations of social media. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, 1–13. DOI: https://doi.org/10.1145/3313831.3376424

[87] Joanna C. Yau and Stephanie M. Reich. 2018. "It's just a lot of work": Adolescents' self-presentation norms and practices on Facebook and Instagram. *Journal of Research on Adolescence*, 29, 1, 196–209. DOI: https://doi.org/10.1111/jora.12376

[88] Xuan Zhao, Cliff Lampe, and Nicole B. Ellison. 2016. The social media ecology: User perceptions, strategies and challenges. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*, May 7–12, 2016, San Jose, CA. ACM, New York, NY, 89–100. DOI: https://doi.org/10.1145/2858036.2858333

[89] Raymond Zhong and Sheera Frenkel. 2020. A third of TikTok's U.S. users may be 14 or under, raising safety questions. *The New York Times*. Retrieved from https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html

[90] Chengyan Zhu, Xiaolin Xu, Wei Zhang, Jianmin Chen, and Richard Evans. 2020. How health communication via TikTok makes a difference: A content analysis of TikTok accounts run by Chinese provincial health committees. *International Journal of Environmental Research and Public Health*, 17, 1, 192. DOI: https://doi.org/10.3390/ijerph17010192

Received January 2021; revised April 2021; accepted July 2021.

PX0325

Help Center

Using Facebook Friending

People You May Know

What is People You May Know on Facebook?

People You May Know is a list of people who you might want to be friends with because you have things in common, like a mutual friend, school or workplace.

You can send a friend request to anyone on your People You May Know list by selecting Add Friend. If you think a suggestion is unhelpful, you can remove it and see fewer suggestions like it.

Was this helpful?

Yes No

How can I manage information used for People You May Know?

There are different ways you can control what information People You May Know uses.

You can:

- Change the privacy settings of your friends list.
- Limit who can send you friend requests to friends of friends.
- Manage mobile phone and email contacts you uploaded to Facebook or Messenger.
- Control who Facebook can suggest your profile to based on your email address or phone number.

Learn more about what information we collect and how it's used in our Privacy Policy.

Was this helpful?

Yes No

Where can I see People You May Know on Facebook?

Help Center

To see People You May Know:

1. From your Feed, click Friends in the left menu. You may have to click **See More** first.
2. Scroll down to **People You May Know** or click **Suggestions** in the left menu.

You may also see People You May Know suggestions in your notifications and Feed.

Was this helpful?

Yes No

Where People You May Know suggestions on Facebook come from

People You May Know can help you find other friends on Facebook.

Friend suggestions come from things like:

- Having friends in common
- Your networks, such as your current city, school or work
- Being in the same Facebook group
- Being tagged in the same photo or post
- Contacts you uploaded

Learn how to manage what information is used for People You May Know.

To add someone as a friend from People You May Know, you can send them a friend request. If you don't like a suggestion then you can remove it. Learn how to hide People You May Know suggestions from your Feed.

Please read our notice on contact uploading for Non-Users.

Was this helpful?

Yes No

How does Facebook use my information to show suggestions in People You May Know?

Help Center

People You May Know suggestions can be friends of friends, people in your network or people you may have something in common with.

What information does Facebook use to show suggestions in People You May Know?

People You May Know suggests people you might be likely to add as a friend on Facebook. Friend suggestions come from things like:

- Having friends in common.
- Your profile information and networks (example: your current city, school or work).
- Your Facebook activity (example: joining groups or being tagged in the same photo or post).
- Contacts you or someone you may know uploaded to Facebook or Messenger.

If you upload your contacts through the Facebook app, learn how to manage contact uploading. If someone has uploaded their contacts and it includes your mobile phone number or email address, you may show up as friend suggestions for each other.

For more information, please see our notice on contact uploading for Non-Users.

Was this helpful?

Yes No

How can I hide People You May Know suggestions from my Facebook News Feed?

You can hide People You May Know from your News Feed temporarily, but you may still see it in other places on Facebook.

To temporarily hide People You May Know from your News Feed:

1. Next to **People You May Know**, click •••.
2. Click **Hide People You May Know**

You can't permanently hide People You May Know from your News Feed.

Was this helpful?

Help Center

I'm seeing people I don't know suggested as People You May Know on Facebook.

People You May Know suggestions come from things like having mutual friends or being a part of the same network. We regularly update People You May Know to improve suggestions. However, sometimes we might get it wrong and show someone you don't know or want to add as a friend.

How do I remove a friend suggestion?

If you think a suggestion is unhelpful, you can remove it. Removing unhelpful suggestions helps improve People You May Know so that you see fewer suggestions like them.

To remove a suggested friend:

- From Friends, click **Remove** next to their name.
- From your News Feed, click ✕ on the picture of the person you want to remove.

Blocking someone stops them from showing up as a friend suggestion, and you also don't show as a suggestion for them. Learn more about what happens when you block someone on Facebook.

You can also temporarily hide People You May Know from your News Feed.

Was this helpful?

Yes No

Information for people who don't use Meta Products

Meta Platforms, Inc. ("**Meta**", "**we**", "**our**" or "**us**") processes your name, mobile phone number and/or your email address if we receive it from our users through the contact uploading or contact syncing feature available on Facebook, Messenger or Instagram ("**Contact Uploading**"). We process this information even if you are not a user of Facebook, Messenger or Instagram and/or don't have an account with us (a "**Non-User**").

About Contact Uploading and how it works

When a user uses Contact Uploading and grants us access to their device's address book. We will access and upload the names, phone numbers and email addresses in their address book on a daily basis to our servers, including those of both users of Facebook, Messenger and/or Instagram and other contacts who are not users or don't have an account (i.e., Non-Users).

Help Center

How do we use a Non-User's information?

Contact Uploading is an optional feature where users can choose to upload their device's address book to Facebook, Messenger and/or Instagram. We process the names, phone numbers and/or e-mail addresses in the address book to see if any of the numbers or email addresses belong to users. If a user's contact is also on Facebook, Messenger and/or Instagram, we can suggest this contact on Facebook as someone to send a friend request to in the Facebook user's People You May Know or as a suggested account to follow on Instagram. Users can also invite Non-Users to join Facebook, Messenger and/or Instagram via text or email.

Contact Uploading also allows us to suggest a Non-User who joins as a user of Facebook as a friend to add for the user through People You May Know or as a suggested account to follow on Instagram.

Finally, we will also use the address books uploaded to investigate suspicious activity on the Meta Products and keep our platform safe and secure. We also conduct business intelligence and analytics using the address books uploaded to accurately count people and users of the Meta Company Products.

Meta shares information we collect, infrastructure, systems and technology with the other Meta Companies. We share Non-Users' data with other Meta Companies to be able to provide the contact import feature and for the purposes set out above.

We retain Non-User's personal information for as long as needed to provide the Contact Importer feature to users and for the purposes described above. In certain cases, we need to keep your contact information for longer, including after you ask us to erase it. This includes for legal reasons such as to respond to a legal request or comply with applicable law, for regulatory or litigation matters, to prevent harm or for safety, security and integrity purposes. The length of time we need to keep it will depend on the specific reason.

How Non-Users can exercise their rights

We provide Non-users the ability to exercise their rights under applicable laws. To exercise your rights, contact us using the contact information below.

You can learn more about how privacy works on Facebook and on Instagram, and in the Facebook Help Center.

We hope that we can satisfy any queries a Non-User may have about the way in which we process their personal information. In some countries you may also be able to contact the Data Protection Officer for Meta Platforms, Inc., and depending on your jurisdiction, you may also contact your local Data Protection Authority ("DPA") directly.

Click here if you have a question about the rights you may have.

How we Transfer information

Help Center

variety of reasons, including so we can operate and provide the services described in this Data Notice.

Non-User information will be transferred or transmitted to, or stored and processed in:

- Places we have infrastructure or data centers, including the United States, Ireland, Denmark and Sweden, among others
- Countries where Meta Products are available
- Other countries where our partners, service providers and vendors are located outside of the country where you live, for purposes as described in this Data Notice.

We rely on appropriate mechanisms for international data transfers. For example, for information we collect:

- We utilize standard contractual clauses approved by the European Commission and by other relevant authorities.
- We rely on determinations from the European Commission, and from other relevant authorities, about whether other countries have adequate levels of data protection.
- We use equivalent mechanisms under applicable laws that apply to data transfers to the United States and other relevant countries.

United States Regional Privacy Notice

If you are a United States resident, you can learn more about your consumer privacy rights by reviewing the United States Regional Privacy Notice.

Contact us

If you have questions about this Notice, or have questions, complaints or requests regarding your information, please use this form.

Alternatively, if you don't have access to the form, you can also write to us at the following address:

Meta Platforms, Inc.

ATTN: Privacy Operations

1601 Willow Road

Menlo Park, CA 94025

Was this helpful?

Yes No

Help Center

English (US)

About

Privacy

Terms and Policies

Ad Choices

Careers

Cookies

Create Ad

Create Page

from Meta

© 2024 Meta

PX0326

Meta Business Help Center

Use location targeting

Reach people based on locations such as country, region or city.

Before you begin

Set up a campaign before selecting your audience.

Open **Meta Ads Manager** and select a campaign objective for your ad set.

Go to **Audience** and select your source audience.

Go to **Locations** and follow the prompts.

Choose your locations

Complete accuracy is not guaranteed with location targeting. Learn more about location targeting.

Sometimes people see an ad organically. In these instances, you may have engagement such as likes, comments and shares on the ad outside of the targeted location. We may also show preview ads to people connected to the business, even if they're outside the targeted location. You'll not be charged for preview impressions.

The location targeting selection you choose will be used to find people living in or recently in the selected location. This helps you reach many people within a specific area during a certain timeframe.

About Audience Location Status

Enter a location

Enter the names of countries, cities and regions in the box below the dropdown. For best results, enter one location at a time. If you want to reach multiple locations, we recommend you bulk upload them.

Locations you can search for

- Countries (up to 25)
- States
- Provinces
- Cities (up to 250)
- Congressional districts
- ZIP or post codes (up to 50,000)

Depending on your campaign objective and audience selections, you can include broad locations, such as **Worldwide** and **Asia**, or geographical regions.

You will reach people living in or recently in the location.

Browse locations

If you don't have a location or aren't sure of one, select **Browse** for suggested countries and regions, or to use your saved locations. You can also enter a country name and get suggestions for high population cities.

Choose a location radius

The radius pindrop is a visual indicator below the location box. It identifies relevant areas inside a larger geographic area or extends your area's reach beyond its geographical boundary. You can either select **Drop Pin** and adjust the area by zooming, or you can select the dropdown from an area listed within the location box and use the slider to adjust the distance.

Note:

- Radius minimum and maximum vary by location type, and don't extend into other countries, even when the visual map extends beyond a border. You need to include each country separately.
- Your audience reach includes all other locations that fall within the radius.
- Radius selection isn't possible when using ZIP or postal codes, states, regions or countries. In those cases, the radius would include an entire area.
- Choose **Current city only** to just advertise in your chosen city.

Learn more

- About reaching new audiences
- About detailed targeting
- About international location targeting
- Countries included in global targeting

PX0333

Policy advisory opinion on Meta's cross-check program

I. Executive summary 3
II. Request from Meta 5
III. Meta's cross-check system 7
Meta's explanation of why it uses cross-check 7
How cross-check works 9
Early Response Secondary Review (ERSR) 10
General Secondary Review (GSR) 18
Cross-check and reported exemptions from enforcement 22
IV. Framework for Board analysis 23
International human rights standards 23
Meta's values 24
V. Assessment of cross-check system 25
Broad scope to serve multiple and contradictory objectives that enables visibility for violating content 26
Unequal access to discretionary policies and enforcement 29
Program enrollment exceeds capacity 31
Failure to track core metrics to assess the program and make improvements 33
Lack of transparency and auditability of the program and its functioning 34
Conclusions about cross check 34
VI. Enforcement recommendations 35
Entity-based mistake-prevention system governance recommendations 36
Users that should be included in entity-based mistake prevention systems 36
Decision makers should be qualified and empowered to make rights-respecting decisions 37
Guidance to create and govern lists for entity-based mistake-prevention systems 38
Guidance to maintain and audit lists for entity-based mistake prevention systems 39
Some entities receiving additional protection should be publicly marked 40
Content-based mistake-prevention system governance recommendations 41
Content that should be selected and prioritized for content-based mistake prevention systems 41
Technical corrections 42
General mistake-prevention system governance recommendations 43
Harm mitigation following identification of violating content 43
Ensuring appeal availability 44
Learning and improvement 45
VII. Transparency recommendations 46

I. Executive summary

In October 2021, following disclosures about Meta's cross-check program in the Wall Street Journal, the Oversight Board accepted a request from the company to review cross-check and make recommendations for how it could be improved. This policy advisory opinion is our response to this request. It analyzes cross-check in light of Meta's human rights commitments and stated values, raising important questions around how Meta treats its most powerful users. As the Board began to study this policy advisory opinion, Meta shared that, at the time, it was performing about 100 million enforcement attempts on content every day. At this volume, even if Meta were able to make content decisions with 99% accuracy, it would still make one million mistakes a day. In this respect, while a content review system should treat all users fairly, the cross-check program responds to broader challenges in moderating immense volumes of content.

According to Meta, making decisions about content at this scale means that it sometimes mistakenly removes content that does not violate its policies. The cross-check program aims to address this by providing additional layers of human review for certain posts initially identified as breaking its rules. When users on Meta's cross-check lists post such content, it is not immediately removed as it would be for most people, but is left up, pending further human review. Meta refers to this type of cross-check as "Early Response Secondary Review" (ERSR). In late 2021, Meta broadened cross-check to include certain posts flagged for further review based on the content itself, rather than the identity of the person who posted it. Meta refers to this type of cross-check as "General Secondary Review" (GSR).

In our review, we found several shortcomings in Meta's cross-check program. While Meta told the Board that cross-check aims to advance Meta's human rights commitments, we found that the program appears more directly structured to satisfy business concerns. The Board understands that Meta is a business, but by providing extra protection to certain users selected largely according to business interests, cross-check allows content which would otherwise be removed quickly to remain up for a longer period, potentially causing harm. We also found that Meta has failed to track data on whether cross-check results in more accurate decisions, and we expressed concern about the lack of transparency around the program.

In response, the Board made several recommendations to Meta. Any mistake-prevention system should prioritize expression which is important for human rights, including expression of public importance. As Meta moves towards improving its processes for all users, the company should take steps to mitigate the harm caused by content left up during additional review, and radically increase transparency around its systems.

Key findings

The Board recognizes that the volume and complexity of content posted on Facebook and Instagram pose challenges for building systems that uphold Meta's human rights

commitments. However, in its current form, cross-check is flawed in key areas which the company must address:

Unequal treatment of users. Cross-check grants certain users greater protection than others. If a post from a user on Meta's cross-check lists is identified as violating the company's rules, it remains on the platform pending further review. Meta then applies its full range of policies, including exceptions and context-specific provisions, to the post, likely increasing its chances of remaining on the platform. Ordinary users, by contrast, are much less likely to have their content reach reviewers who can apply the full range of Meta's rules. This unequal treatment is particularly concerning given the lack of transparent criteria for Meta's cross-check lists. While there are clear criteria for including business partners and government leaders, users whose content is likely to be important from a human rights perspective, such as journalists and civil society organizations, have less clear paths to access the program.

Delayed removal of violating content. When content from users on Meta's cross-check lists is identified as breaking Meta's rules and while undergoing additional review, it remains fully accessible on the platform. Meta told the Board, that, on average, it can take more than five days to reach a decision on content from users on its cross-check lists. This means that, because of cross-check, content identified as breaking Meta's rules is left up on Facebook and Instagram when it is most viral and could cause harm. As the volume of content selected for cross-check may exceed Meta's review capacity, the program has operated with a backlog which delays decisions.

Failure to track core metrics. The metrics that Meta currently uses to measure cross-check's effectiveness do not capture all key concerns. For example, Meta did not provide the Board with information showing it tracks whether its decisions through cross-check are more or less accurate than through its normal quality control mechanisms. Without this, it is difficult to know whether the program is meeting its core objectives of producing correct content moderation decisions, or to measure whether cross-check provides an avenue for Meta to deviate from its policies.

Lack of transparency around how cross-check works. The Board is also concerned about the limited information Meta has provided to the public and its users about cross-check. Currently, Meta does not inform users that they are on cross-check lists and does not publicly share its procedures for creating and auditing these lists. It is unclear, for example, whether entities that continuously post violating content are kept on cross-check lists based on their profile. This lack of transparency impedes the Board and the public from understanding the full consequences of the program.

The Oversight Board's recommendations

To comply with Meta's human rights commitments and address these problems, a program that corrects the most high-impact errors on Facebook and Instagram should be structured substantially differently. The Board has made 32 recommendations in this area, many of which are summarized below.

As Meta seeks to improve its content moderation for all users, it should prioritize expression that is important for human rights, including expression which is of special public importance. Users that are likely to produce this kind of expression should be prioritized for inclusion in lists of entities receiving additional review above Meta's business partners. Posts from these users should be reviewed in a separate workflow, so they do not compete with Meta's business partners for limited resources. While the number of followers can indicate public interest in a user's expression, a user's celebrity or follower count should not be the sole criterion for receiving additional protection. If users included due to their commercial importance frequently post violating content, they should no longer benefit from special protection.

Radically increase transparency around cross-check and how it operates. Meta should measure, audit, and publish key metrics around its cross-check program so it can tell whether the program is working effectively. The company should set out clear, public criteria for inclusion in its cross-check lists, and users who meet these criteria should be able to apply to be added to them. Some categories of entities protected by cross-check, including state actors, political candidates and business partners, should also have their accounts publicly marked. This will allow the public to hold privileged users accountable for whether protected entities are upholding their commitment to follow the rules. In addition, as around a third of content in Meta's cross-check system could not be escalated to the Board as of May-June 2022, Meta must ensure that cross-checked content, and all other content covered by our governing documents, can be appealed to the Board.

Reduce harm caused by content left up during enhanced review. Content identified as violating during Meta's first assessment that is high severity should be removed or hidden while further review is taking place. Such content should not be allowed to remain on the platform accruing views simply because the person who posted it is a business partner or celebrity. To ensure that decisions are taken as quickly as possible, Meta should invest the resources necessary to match its review capacity to the content it identifies as requiring additional review.

II. Request from Meta

1. The Oversight Board first became aware of cross-check in 2021 when deciding its case on the suspension of former US President Donald Trump's accounts. Although Meta did not mention cross-check in its initial referral or materials sent to the Board, it described the cross-check program in response to a Board question about any different treatment the account may have received. As part of its May 2021 decision, the Board made two recommendations relevant to the cross-check program:

- "Produce more information to help users understand and evaluate the process and criteria for applying the newsworthiness allowance, including how it applies to influential accounts."

- "The company should also clearly explain the rationale, standards and processes of the cross-check review, and report on the relative error rates of determinations made through cross-check compared with ordinary enforcement procedures."

2. In September 2021, the Wall Street Journal revealed documentation produced by former employee and company critic Frances Haugen. The Journal's reporting described cross-check as exempting Meta's most influential users from normal content moderation processes. The Independent reported that Frances Haugen said the company had "lied" to the Board about cross-check "repeatedly" during the Trump case. Internal Meta documentation published by the Journal revealed that some of its employees considered cross-check's 'whitelisting' practices "not publicly defensible." Similarly, according to the Journal, users benefiting from the cross-check system at the time were given a 24-hour "self-remediation" window to edit or remove violating content and thus avoid any Meta-imposed penalties.

3. On September 21, 2021, following the Wall Street Journal articles, the Board called on Meta to commit to transparency about the system. The following day, Meta held a briefing with the Board on cross-check. The Board concluded that "the team within Facebook tasked to provide information has not been fully forthcoming in its responses on cross-check. On some occasions, Facebook failed to provide relevant information to the Board, while in other instances, the information it did provide was incomplete."

4. Shortly after the Board called for greater transparency on cross-check, Meta submitted this policy advisory opinion request. After briefly summarizing the system, Meta described cross-check as a program that "provides additional levels of review for certain content that our internal systems flag as violating (via automation or human review), with the goal of preventing or minimizing the highest-risk false-positive moderation errors." Meta defines false positives as the mistaken removal of content that does not violate the content policies that establish what is allowed on Facebook and Instagram.

5. Meta posed the following three questions to the Board:

   Because of the complexities of content moderation at scale, how should Facebook balance its desire to fairly and objectively apply our Community Standards with our need for flexibility, nuance, and context-specific decisions within cross-check?

   What improvements should Facebook make to how we govern our Early Response ("ER") Secondary Review cross-check system to fairly enforce our Community Standards while minimizing the potential for over-enforcement, retaining business flexibility, and promoting transparency in the review process?

   What criteria should Facebook use to determine who is included in ER Secondary Review and prioritized as one of the many factors by our cross-check ranker in order to help ensure equity in access to this system and its implementation?

6. The Board accepted Meta's request on October 21, 2021. Following this acceptance, the Board sent Meta questions. The Board asked Meta 74 questions. 58 were answered fully, 11 were answered partially, and five were not answered. Meta took months to respond to some of these questions.

7. The Board also received 87 public comments related to this policy advisory opinion: nine from Asia Pacific and Oceania, two from Central and South Asia, 12 from Europe, three from Latin America and the Caribbean, three from the Middle East and North Africa, three from Sub-Saharan Africa, and 55 from the United States and Canada. To read public comments submitted for this policy advisory opinion please click here. In addition, the Board held four regional workshops focused on the cross-check program.

8. Based on its analysis of this information, independent research, and stakeholder engagement, the Board now answers Meta's questions and provides its assessment of the cross-check system. Meta also told the Board it has made significant changes to the cross-check program over the past year. The Board understands these changes to be, at least in part, an effort to respond to public criticisms of the program. The Board's explanation of the program and its analysis of it is based on how Meta states the program is currently functioning. However, at times the Board references its understanding of past practices as they inform likely areas of recurrent risk.

9. The Board explored whether the program serves in practice to address and mitigate adverse impacts according to Meta's human rights responsibilities. This analysis, grounded in international human rights standards and Meta's stated values and commitments, implicates important questions of how Meta treats its most influential and powerful users, permits content to flow across its platforms, and provides information to the public about its actions.

III. Meta's cross-check system

Meta's explanation of why it uses cross-check

10. Facebook and Instagram users create billions of pieces of content each day. Meta is constantly moderating content; or screening, evaluating, and taking action on it based on the company's content policies. On Facebook, these policies are the Community Standards, and on Instagram, they are the Community Guidelines.

11. According to Meta, moderating content at this scale presents challenges, and its human reviewers and automated systems sometimes mistakenly remove content that does not violate Meta policies. Meta refers to these decisions as false positives. False negatives are a form of under-enforcement and refer to content that violates Meta policies but is not determined to be violating on review. Under-enforcement also includes violating content that is not detected by automated or human

reviewers, and system design choices that allow violating content to remain visible after a first review.

	Review determined content does not violate the Community Standards	Review determined content violates the Community Standards
Content does not violate the Community Standards	True Negative	False positive (over-enforcement)
Content violates the Community Standards	False negative (under-enforcement)	True positive

12. The cross-check system only addresses over-enforcement, or false positives. Through this system, Meta delays taking any enforcement action on select content initially identified as violating to allow for possible additional review with the aim of avoiding false positives.

13. Meta described cross-check as a mistake-prevention strategy that allows it to balance protecting users' voice from false positives with the need to quickly remove violating content. As part of the policy advisory opinion request, Meta highlighted the inclusion of "journalists reporting from conflict zones and community leaders raising awareness of instances of hate or violence," as well as civic actors where "users have a heightened interest in seeing what their leaders are saying."

14. The system further includes users that Meta describes as "business partners." These partners have dedicated points of contact at Meta. According to the company, these users include "health organizations, news publishers, entertainers, musicians, artists, creators and charitable organizations." The Board understands that this category includes users that are likely to generate money for the company, either through formal business relationships or because they draw users to the platform and keep them engaged there. The Board understands that "business partners" likely also include major companies, political parties and campaigns, and celebrities.

15. Meta told the Board that it adds "business partners" to cross-check to prevent mistaken deletions that limit the ability of users and advertisers to reach their audience and customers, and the economic and reputational impact that such errors may cause the company. For these users, Meta aims to avoid "negative experiences for both Facebook's business partners and the significant number of users who follow them."

16. Meta stated that it prefers under-enforcement compared to over-enforcement of cross-checked content, as “in the current business landscape maximizing the benefit of cross-check (preventing false positives) is generally considered to be more important than minimizing the cost of cross-check [i.e., views of violating content]. This is due to the perception of censorship.” The Board interprets this to mean that, for business reasons, addressing the “perception of censorship” may take priority over other human rights responsibilities relevant for content moderation.

How cross-check works

17. Meta’s ordinary content moderation processes apply to most users. When content is identified as violating Meta’s content policies, Meta takes an enforcement action. This includes content deletion and the application of warning screens, depending on the type of policy violation. Some violations can also lead to account-level penalties, such as suspension and termination. However, in some cases, content receives a different treatment, as is the case of the cross-check system.

18. Meta uses the term cross-check to refer to a false positive-prevention program. It provides for additional layers of review for content before enforcement action is taken. Escalation-only content policies, which can only be applied by specialized teams at Meta, may be applied during this enhanced review. These policies include the newsworthiness and spirit of policy allowances and all rules that Meta has determined require additional context to enforce. Cross-check review processes are triggered under two sets of circumstances.

19. First, cross-check provides guaranteed additional **human review of content** by specific entitled entities whenever they post content that is identified as requiring enforcement under Meta content policies. Meta calls this **Early Response Secondary Review** or **ERSR**. An "entity” is anything on Facebook or Instagram that can post content, such as Facebook pages, Facebook profiles, and Instagram accounts. Entities can represent individual people and groups or organizations. Meta creates and maintains lists of entities it has decided are entitled to receive the benefits ERSR provides. This means that if any entitled entity posts content that is identified as violating the Community Standards or Guidelines, it will not be removed according to the procedures that apply to regular users, but instead will be sent for extra levels of review. Because ERSR is based on lists, only certain pre-selected users receive this benefit.

20. The second part of the cross-check system provides additional review of certain content identified as violating Meta policies, regardless of the identity of the user who posted it. Meta calls this **General Secondary Review**, or **GSR**. Whenever any piece of content posted by any entity on the platform is identified as violating a Meta policy, by a human or automation, Meta uses an automated process called ‘cross-check ranker’ to instantaneously analyze various factors and determine if the content should be sent for additional review, and how it should be prioritized within

a queue of other content awaiting the same type of review. According to Meta, because this system is based on the characteristics of the content, content posted by any user on Facebook or Instagram is eligible to be selected for GSR. GSR was implemented in 2021, and the Board understands that, to some extent, it was developed and implemented across the platform in response to criticism of ERSR, including the Haugen revelations.

21. The initial detection of content in both types of cross-check that may trigger a review can happen either proactively, through automated Meta systems after the content is posted, or reactively, following user reports. The enforcement actions that may trigger cross-check review include content deletion and the application of warning screens, depending on the type of policy violation. As most content policy violations can lead to account-level penalties, such as suspension and termination, these types of enforcement are also impacted. Cross-check applies across Facebook and Instagram, except for some content types (e.g., reels, podcasts) that are not currently eligible for the program. According to Meta, "10% of organic content that is otherwise subject to integrity enforcement is not eligible for cross-check review today."

22. During the time after content eligible for cross-check (through GSR or ERSR) is identified for enforcement but before it is subject to the additional review process, this content remains fully accessible on the platform, even if the first assessment is that the content violates the Community Standards or Guidelines.

23. The Board understands that if Meta had more moderators available, more content in cross-check review queues would receive an additional human review. However, Meta has chosen to only guarantee additional human review for content that goes through ERSR, the system for entitled entities. Meta has not invested the resources needed for all content identified through GSR to receive additional human review. While the review paths for these two mechanisms differ, as described below, should any reviewer at any stage of the process find that content does not violate Meta policies, the review process ends, and the content remains on the platform.

Early Response Secondary Review (ERSR)

24. Meta states that it includes entities on ERSR lists by assigning them a "tag" that correlates to the nature and sensitivity of the entity. Specific tags correspond to different ERSR lists. Meta states that it applies an ERSR tag to entities corresponding to the following categories: (1) civic and government; (2) significant world events; (3) media organizations, businesses, communities and creators, including advertisers; (4) historically over-enforced; (5) legal and regulatory or entities for which erroneous action may present legal risk to Meta, for example in the context of ongoing litigation; (6) entities whose content is under review, meaning cases where action by any reviewer would undermine ongoing deliberation or would present risk to Meta. According to Meta, beyond the factors that it uses to determine whether an entity fits within any of the categories stated above, such as advertisement

spending or history of enforcement, entitlement to ERSR is also determined by an assessment of the impact a potential enforcement mistake would have on the company in terms of the level of company leadership that would be involved in finding a solution. In other words, a key rationale for ERSR is to avoid provoking people who have the means to engage senior-level executives directly or create public controversy those executives might need to remedy.

25. Meta informed the Board that it is currently consolidating and updating its ERSR lists. Previously, Meta's lists corresponded to the level of escalation that would be required to enforce content policies against a particular entity. According to Meta, all entities currently entitled to ERSR are now subject to the same review process. This process may include discretionary escalation to the highest levels of the company.

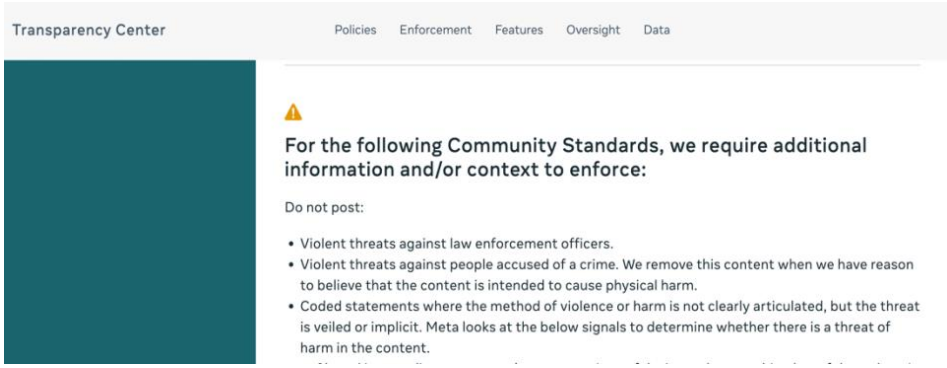
26. Meta told the Board that during the second quarter of 2022 it established general criteria for adding and removing entities from ERSR lists, and new processes for periodic audits and internal oversight. Meta did not provide details on these processes, and what actions might trigger the re-evaluation and removal of an entity. Meta did explain that, in general, the tags that place an entity on an ERSR list will expire after a year, and in theory the entitled entities would need to be assessed and tagged anew. According to Meta, this logic generally covers entities in the following categories: legal and regulatory; significant world events; media organizations; businesses, communities, and creators; historically over-enforced; and entities escalated for higher context review. Meta noted two exceptions to the one-year expiration rule. First, tags for entities in the civic and government category do not have a default expiration. Second, tags for entities in the other categories mentioned above may be given shorter ERSR entitlement at Meta's discretion.

27. Whenever a piece of content by any of the entitled entities is marked for enforcement by automated or human review, no enforcement action is taken, and the content is instead sent for **enhanced review by a human moderator**. This first level of enhanced review is done by what Meta refers to as a "**Regional Market Team**," a team within Meta. This team includes both Meta employees and hired contractors who have additional contextual and language knowledge about a specific geographic market. If a Market Team reviewer determines the content is non-violating, the process ends, and the content remains on the platform.

28. However, if the Market Team reviewer finds the content violates Meta's policies, the content remains on the platform while it is escalated further to what Meta calls the "**Early Response Team**" for another review. According to Meta, this team has "deeper policy expertise and the ability to factor in additional context."

29. The Early Response Team is also allowed wider discretion than other Meta content moderators and can apply content policies that "require additional information or context to be enforced." Meta often marks these content policies with a yellow exclamation point within each Community Standard, as shown below. For example,

at the end of Facebook's Violence and Incitement Community Standard, Meta prohibits "violent threats against law enforcement officials." According to Meta, the determination of whether to keep up or remove content that may violate these context-specific parts of the policy may only be made by a team that is allowed to factor in additional context, like the **"Early Response Team**."



30. The **Early Response Team** may also apply what Meta calls its "newsworthiness" and "spirit of the policy" allowances, which allow otherwise violating content to remain on the platform because Meta finds it is in the public interest or finds that, even though it violates the letter of a policy, it does not violate the intent of the policy. The Board also believes this discretion extends to the application of account-level penalties. However, as disclosed by Meta, the **Early Response Team** does not have language or regional expertise and it relies on translations and contextual information provided by the relevant Regional Market Team to assess the content.

31. At the time of the Board's briefings with Meta, approximately 0.01% of all content identified as needing enforcement under a Meta policy was escalated through cross-check to reviewers who may apply these contextual policies and allowances. Content posted by users on ERSR lists is guaranteed to reach those reviewers before any enforcement action: it may not be removed or have a warning screen applied by automated review, at-scale human reviewers, or **Market Team** reviewers. During the entire time the cross-checked content is awaiting its final determination, it remains on the platform, where users are free to like and share it.

32. Once the content is reviewed by the **Early Response Team**, if it is found violating, Meta may take the corresponding enforcement action, such as removing the content or applying a warning screen. However, Meta may also escalate the decision further. The Board understands that the escalation procedures at this phase are broadly discretionary. If the Early Response Team finds that the content "is an extreme edge case interpretation of [Meta's] policies" or it "presents a significant risk to the company or the community, and/or there is disagreement among internal stakeholders on how to respond," the Early Response Team may perform an additional review in conjunction with other Meta teams. According to Meta, these escalated reviews "include the views and input of the Content Policy subject matter experts (SMEs) and the local Public Policy, Comms, and Legal teams"

and may include input from other teams. After that review, it could even be escalated further to company leadership before receiving any enforcement action.

33. Additionally, Meta told the Board that if "the issue has significant service blocking, legal, regulatory, or safety risk, or where [it has] limited time to make a decision, [the Early Response Team] will on rare occasions escalate a decision directly to global senior leadership." Meta stated it assesses liability risk, urgency, geopolitical impact, service blocking risk and disagreement between internal teams as factors to escalate these decisions.

34. In summary, a piece of content posted by an entity on an ERSR list may receive up to five reviews before it is subject to enforcement, even if reviewers repeatedly find that it violates Facebook or Instagram rules and escalate it along the cross-check pathways:

1. Initial review by automation or a human reviewer that identifies content for enforcement based on Meta's policies.
2. Regional Market Team review.
3. Early Response Team review. This is the first review that can authorize enforcement against the content. This team may request an enhanced Early Response Team review that incorporates other teams or pass directly to Global Leadership review.
4. Enhanced Early Response Team review with subject matter experts, Public Policy, Communications and Legal Teams.
5. Global Leadership review. This is a discretionary escalation from the Early Response Team based on the severity of consequences to the company.

The process stops and the content remains on the platform if it is found non-violating at any stage of review.

How Cross-Check Works

ERSR: Early Response Secondary Review

Available only for "entitled entities" within special categories.
Intended to correct mistaken over-enforcement.

All content

Initial Detection

Initial detection of content that needs to be reviewed can come from two sources:
(1) Automated detection (i.e. when algorithms flag content for review);
(2) Detection by a human when they report content.

First Decision

At- scale or outsourced review

For the vast majority of users a decision is reached by an automated algorithm or an outsourced human reviewer.

But for 'entitled entities'* the cross-check system now kicks in.

If violating →

TECHNICAL CORRECTIONS

This is the practice of allowing certain entities to be exempt from content moderation enforcement actions for specific policies.

Cross-Check

Enhanced Review by Human Moderator

Regional Market Team

These are Meta employees or contractors with additional relevant knowledge, including language and regional expertise.

If violating →

Another Review by Human Moderator

Early Response Team

These are Meta employees with deeper policy expertise. They can apply greater discretion in enforcement decisions. For instance, they can enforce allowances for 'newsworthiness,' or make decisions based on the the 'spirit of a policy' rather than the letter of the Community Standards, where this is in the public interest. They are not required to have language or regional knowledge.

If case is complex →

Panel Review

Enhanced Early Response Review

This is a panel of more senior Meta executives, including their comms, legal and public policy teams.

If decision is not straightforward →

Top Level / Urgent Review

Global Leadership

The most senior Meta executives may be involved if there's significant legal, regulatory or safety risk in a decision.

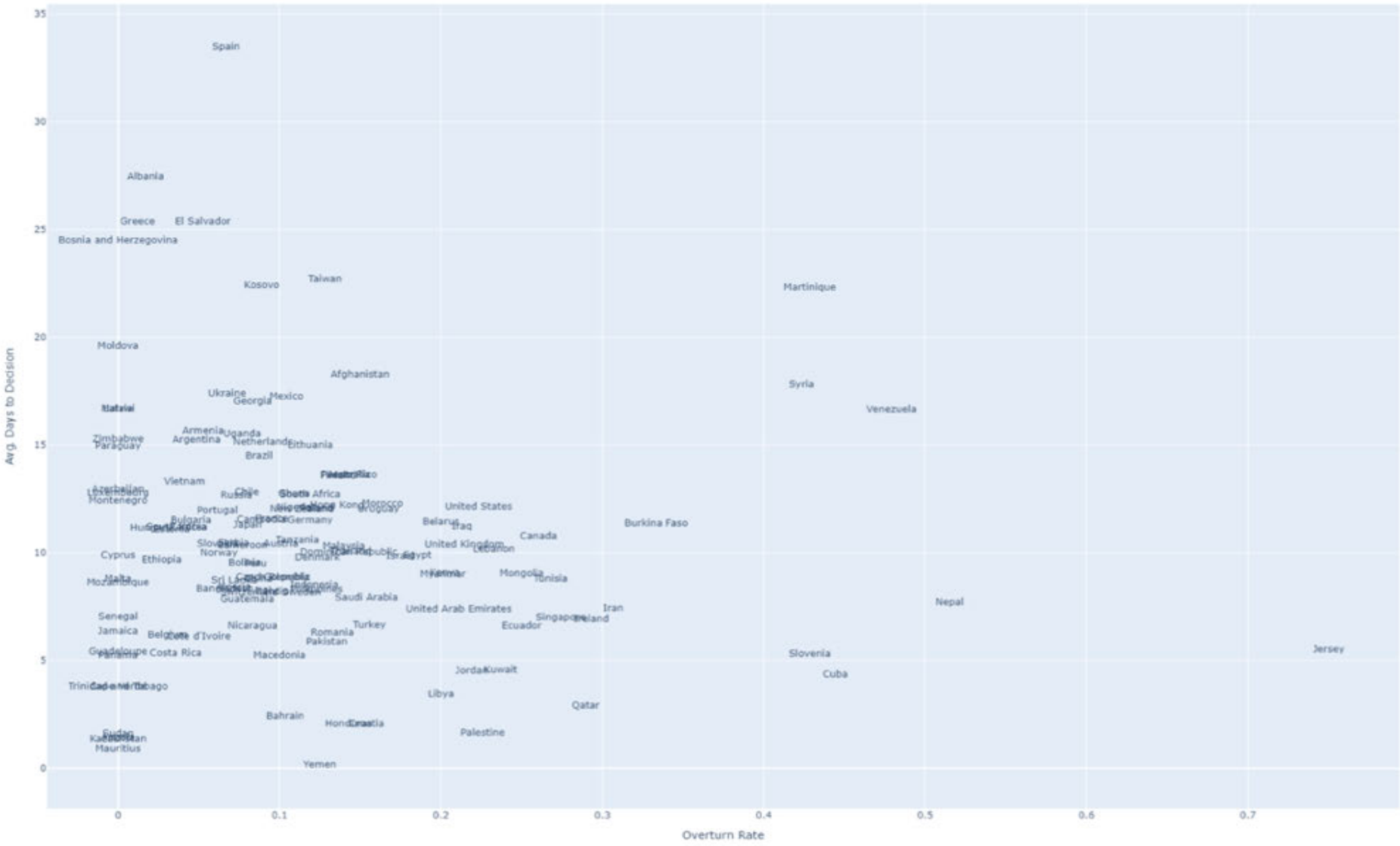
If urgent and with consequences to the company →

***Categories, as defined by Meta, of entitled entities that benefit from Early Response Secondary Review**

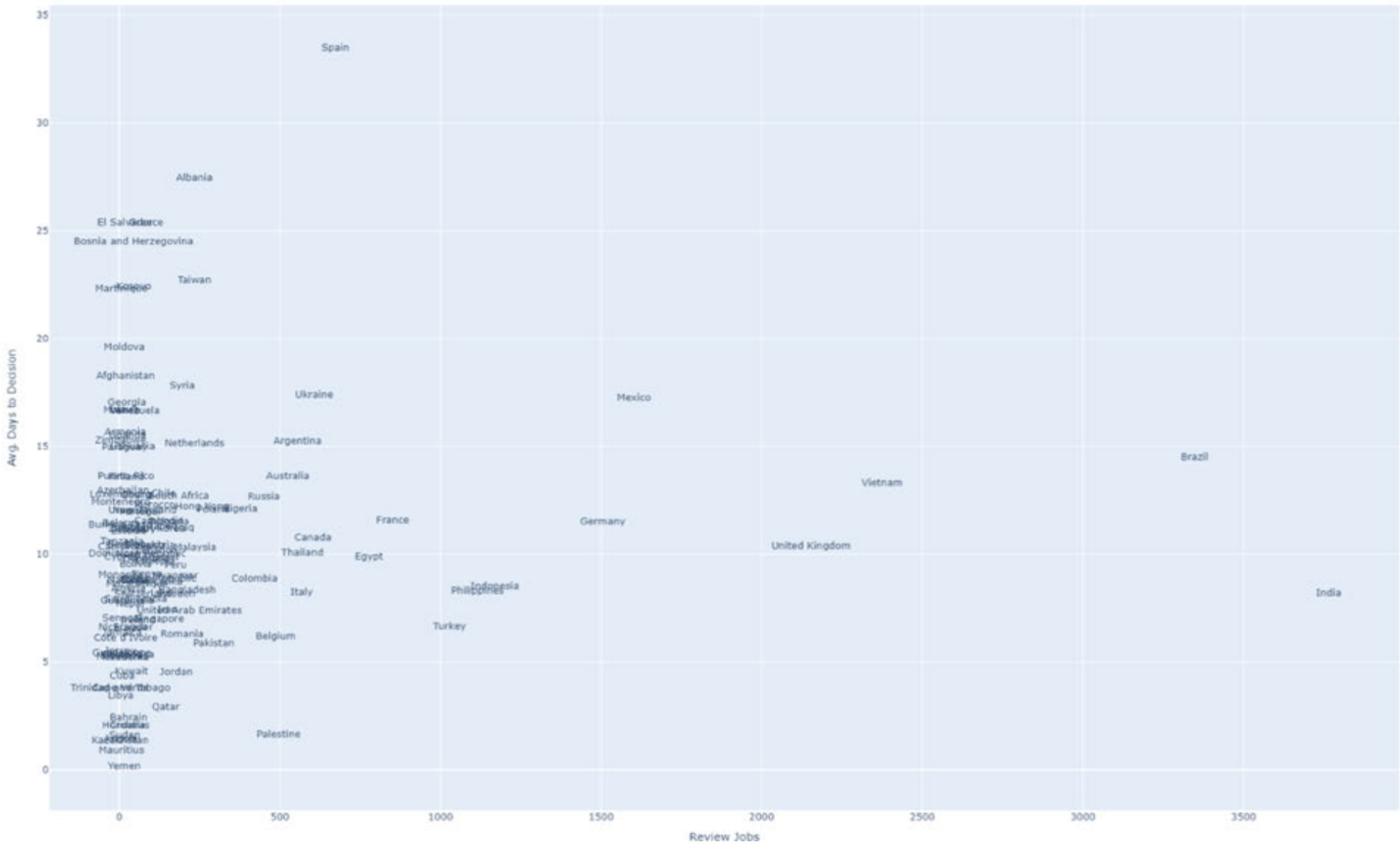
**Entities related to escalation responses or high-risk events*
**Entities included for legal compliance purposes*
**High-visibility public figures and publishers*
**Marginalized populations*
**Civic entities*

Content remains accessible throughout these reviews until the final removal decision is made.

35. The ERSR pathway may take several days. According to Meta, its internal goal for cross-check review by the Market Team ranges from 12 to 120 hours depending on the severity of the potential violation. In practice, Meta stated the mean time to final decision for Early Response Secondary Review is more than five days. For content posted by users in the United States, Meta states that it takes "approximately 12 days on average for a decision." Other countries have even slower mean resolution times. For instance, the mean time-to-decision for Afghanistan and Syria is about 17 days. In the information Meta provided the Board, the longest time a piece of content remained in the Early Response Secondary Review queue was 222 days. Meta provided several graphs with this data to the Board, displaying the mean time to decision between March 2021 and February 2022, with the overturn rate and the number of jobs, or pieces of content reviewed, and different countries.



(1) Scatterplot of average days to decision vs. overturn rate by country



(2) Scatterplot of average days to decision vs. number of jobs by country (excl. USA)

36. Meta states that content that spends more time awaiting review has been designated as low severity under its "violation severity framework." This scheme ranks content based on the specific Community Standard that the first review indicates it violates. Meta's framework ranks each Community Standard according to the potential harm that policy violations may cause, a determination that Meta has said it made based on the company's research. For example, it considers hate speech more harmful than spam, with potential hate speech prioritized ahead of spam in the ERSR queue.

37. That said, in September 2021, the Wall Street Journal reported that Brazilian soccer star Neymar posted non-consensual intimate imagery of another person on his Facebook and Instagram accounts. According to reporting by The Guardian, the video was online for over a day, and "an internal review of the Neymar posts found that the video was viewed 56 million times on Facebook and Instagram before removal" despite representing a clear violation of Meta content policies. According to Meta, the reason for the prolonged accessibility of this violating content was a "delay in reviewing the content due to a backlog at the time."

38. A core metric that Meta told the Board it uses to justify the cross-check system and assess how well it is working is the "overturn rate." This is the percentage of content that is found non-violating during cross-check review, reversing the initial determination and preventing the enforcement of content that Meta's rules allow. Meta provided several different figures to the Board about its overturn rate for ERSR content. According to Meta, for different time periods over the last year the overturn rate ranged from 30% to 90%. When the overturn rate is low, ERSR is keeping more content ultimately found violating on the platform during the multiple layers of cross-check review. When the overturn rate is high, ERSR is preserving more non-violating content from mistaken enforcement.

39. According to Meta, "most views happen when content is fresh, so speed in reviewing decisions and removing content quickly is crucial in preventing harm." Therefore, violating content that is subject to ERSR remains accessible on the platform throughout the period during which it is likely to receive the vast majority of its views.

General Secondary Review (GSR)

40. The second mechanism that Meta says forms part of its cross-check system is **General Secondary Review** (GSR). Whereas ERSR applies to all content posted by specific entitled entities, GSR may apply to any content posted on the platform, regardless of the poster, based on an algorithmic determination.

41. GSR is a relatively new system. In the fall of 2021, when Facebook whistleblower Frances Haugen disclosed information about cross-check, the Board understands her to have been referring to its previous iteration, which was based entirely on the

entity posting the content. The company now includes GSR within cross-check as well. Meta stated that GSR was available for content posted by some users in 2021 and fully implemented for content posted by all users in early 2022.

42. To decide what content to send through the GSR pathway before taking an enforcement action, Meta uses an algorithm (i.e., a set of rules that need to be followed by a computer for a specific task) called cross-check ranker. This algorithm is based on the following features: "topic sensitivity (how trending/sensitive the topic is), enforcement severity (the severity of the potential enforcement action), false positive probability, predicted reach, and entity sensitivity." Entity sensitivity is therefore a factor in both systems, although in ERSR it is the key factor and in GSR it is one factor among many. Meta stated that it has considered including additional factors and expects to do so in the future.

43. According to Meta, content must satisfy two conditions to be eligible for GSR. First, it needs to have been identified for enforcement (i.e., violating a Community Standard or Guideline) by automation or human review. Second, it must be marked by the cross-check ranker as high priority. If both conditions are met, the content is not enforced immediately and instead is placed in a queue for additional human review by a **Regional Market Team**. These are the same Market Teams that also perform the first enhanced review of content posted by ERSR-entitled entities.

44. The Market Teams are unable to review all content that is guaranteed review under ERSR and all the content that is placed in a queue for possible review under GSR. Because entitled entities on ERSR lists are guaranteed review, the Market Teams must first dedicate reviewer capacity to this content. With any remaining capacity, the relevant Market Team reviews the algorithmically identified GSR content. The Market Teams also review certain content outside the cross-check program, among other tasks they must prioritize.

45. Therefore, even though GSR content may be highly prioritized by the cross-check ranker algorithm as meriting additional review because it may have been identified as a likely false positive, the Market Team may not have capacity to review it. In some cases, if there is no review capacity at the Market Team level, and Meta has chosen to make available its outsourcing capacity, some GSR content may be sent for that additional review to outsourced human reviewers. If the GSR content is reviewed by a Market Team reviewer, in most cases that decision is final. If the content is found violating, it is generally subject to enforcement (e.g., removed or warning screen applied). If it is found not violating, it remains on the platform. However, if the **Early Response Team** has any additional capacity after its obligations to review all ERSR content prior to its possible removal, that team may review highly-prioritized GSR content that a Market Team reviewer has found violating before Meta proceeds to enforcement.

46. Similarly to the ERSR process, content that has been originally assessed as violating the Community Standards and placed in the GSR queue remains on the platform

while awaiting additional review. However, unlike ERSR, content in the GSR queue pending review will not remain on the platform indefinitely. Content that is not reviewed eventually "times out" of the GSR queue. When this happens, Meta reverts to its initial enforcement decision without further review. This means the action that would have been applied, such as removal or a warning screen, is applied on a delayed basis without additional review. If reviewers do not reach a specific piece of content in the GSR queue, it will stay on the platform for between two and four days before Meta removes it from the review queue and applies the enforcement action. At the same time, cross-check ranker continually identifies newer and more highly prioritized content, suspends enforcement on the content, and adds it to the GSR queue.

47. The overall effectiveness of GSR is limited by Meta's choice of how much reviewer capacity to provide to this type of review in each of its markets. The majority of GSR content is reviewed by an outsourced reviewer, a Market Team reviewer, or times out of the system. This means most GSR content never reaches the **Early Response Team** and, therefore, will never reach a level of review where contextual analysis, escalation-only policies, and policy allowances can be applied.

48. Meta also calculates the overturn rate for content that receives cross-check review through the GSR pathway. Meta provided the Board with different rates for this figure over the past year. At the time of the Board's briefings with Meta in February 2022, the overturn rate for General Secondary Review was about 80%. Meta later provided new information to the Board, stating that between March 2022 through May 2022, the overturn rate was about 70%. While there was also variation, these figures varied less over time. Most GSR content initially identified as violating is found to not violate any Meta policy with secondary review. As content times out of the GSR queue, it is highly likely that Meta is enforcing a significant number of false positives identified by cross-check ranker.

How Cross-Check Works

GSR: General Secondary Review

Can apply to any piece of content by any user.
Implemented fully in 2022.

All content

Initial Detection

Initial detection of content that needs to be reviewed can come from two sources:
(1) Automated detection (i.e. when algorithms flag content for review);
(2) Detection by a human when they report content.

→

First Decision

At- scale or outsourced review

For the vast majority of users a decision is reached by an automated algorithm or an outsourced human reviewer.

→ *If violating*

TECHNICAL CORRECTIONS

This is the practice of allowing certain entities to be exempt from content moderation enforcement actions for specific policies.

Cross-Check

Cross- Check Ranker

This is an automatic 'triage' process to analyze content & decide if it needs extra review. It also decides priority of review. It considers: topic sensitivity; enforcement severity; false-positive probability; predicted reach; and the nature and importance of the entity.

→ *If it passes the threshold*

Enhanced Review by Human Moderator

Regional Market Team

These are Meta employees or contractors with additional relevant knowledge, including language and regional expertise.

→ *If violating and there is extra capacity*

Another Review by Human Moderator

Early Response Team

These are Meta employees with deeper policy expertise. They can apply greater discretion in enforcement decisions. For instance, they can enforce allowances for 'newsworthiness,' or make decisions based on the the 'spirit of a policy' rather than the letter of the Community Standards, where this is in the public interest. They are not required to have language or regional knowledge.

Content remains accessible throughout these reviews until the final removal decision is made.

Cross-check and reported exemptions from enforcement

49. The Wall Street Journal reporting described cross-check as a system to exempt "VIP users from the company's normal enforcement." Meta disclosed to the Board that it does have a system that blocks some enforcement actions outside of the cross-check system. Meta refers to this practice as "technical corrections," and public reporting has described it as "allowlisting" and "whitelisting."

50. "Technical corrections" are automatic exceptions to content policy enforcement. This means they override almost all automated or human reviewer attempts to apply an enforcement action for a preselected set of content policy violations. Every piece of content identified for enforcement is automatically checked to see if any "technical corrections" apply.

51. If the content is protected by a correction, it will be exempt from that specific enforcement. As explained by Meta, a "technical correction" applies only to a specific entity for a specific policy violation and does not serve to bar enforcement for other policy violations. At the time of the Board's briefings with Meta, it stated that it applied about a thousand **technical corrections** per day. Meta did not disclose how many and what type of entities have benefited from a "technical correction."

52. If the content is not protected by any correction, it is then checked for cross-check eligibility. At that point, Meta's normal cross-check processes to identify if the user is an ERSR entitled entity or if the content is prioritized by cross-checker ranker for the GSR queue apply.

53. Meta first stated that it primarily applies "technical corrections" to "two violation type groups (spam/inauthentic behavior and impersonation)." Meta later confirmed that as of September 21, 2022, there are four active "technical corrections" and that this might also change over time.

54. Meta told the Board that "a limited number of 'technical corrections' remain, and [Meta] recognize[s] an ongoing need for them." According to Meta, such "corrections help [Meta] prevent enforcement mistakes on content or entities that are highly unlikely to violate our policies and direct human review resources where needed most."

55. Meta acknowledged shortcomings about its past technical corrections practices. Meta told the Board that the "lack of governance over practices in the past, [...] inadvertently resulted in some entities not receiving many enforcement actions." Meta stated that "different teams could apply different corrections to the same entity in a way that, when combined, resulted in the entity and its content not receiving a wide variety of enforcement actions." Meta stated that because this practice was "the inadvertent result of a decentralized system, [Meta] [is] taking

steps to ensure that there is a governance structure around the use of cross-check lists."

Cross-check in the context of government requests to remove content

56. When governments request that Meta remove content, Meta may remove the content because it violates company content policies. It also may remove or "geoblock" content for legal reasons, limiting its accessibility in certain areas. Meta has told the Board that it adds entities to cross-check ERSR lists to protect them from erroneous actions that may present legal risk to Meta, for example in the context of ongoing litigation.

57. According to Meta, government requests to remove content are addressed by specialized teams that may enforce on content immediately, regardless of whether it was posted by an ERSR entitled entity or could have been highly prioritized by cross-check ranker. In other words, removals resulting from government requests supersede cross-check privileges.

IV. Framework for Board analysis

International human rights standards

58. On March 16, 2021, Meta announced a Corporate Human Rights Policy, where it outlines its commitment to respect rights in accordance with the UN Guiding Principles on Business and Human Rights (UNGPs). The UNGPs, endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of businesses. These rights include, "at minimum, [...] those expressed in the International Bill of Human Rights," (Principle 12).

59. As a global corporation committed to the UNGPs, Meta should respect international human rights standards wherever it operates and address any adverse human rights impacts (Principle 11). This also means that Meta should "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts" (Principle 13).

60. The UNGPs also establish that businesses should carry out human rights due diligence to assess actual and potential impacts and act upon their findings (Principle 17). To do that effectively, businesses should monitor qualitative and quantitative indicators and incorporate input from impacted stakeholders (Principle 20).

61. Through its cases, the Board assesses the human rights impacts of specific enforcement decisions. When these cases reveal that Meta is causing a negative impact, or may not be taking steps to identify, monitor, and limit negative impacts more broadly, the Board makes appropriate corrective recommendations. In a

policy advisory opinion, the Board focuses directly on Meta's policy choices, including development and enforcement processes, to assess whether the company is upholding its commitment to respect rights under the UNGPs.

62. Applied to cross-check, the Board explored whether the program serves in practice to address and mitigate adverse human rights impacts according to Meta's responsibilities. The Board also closely examined the metrics Meta uses to determine the effectiveness of the program, and what that suggests about the company's objectives.

63. In its analysis, the Board finds that a wide array of rights may be impacted by the cross-check program. Freedom of expression, which includes the right to seek and receive information (Article 19, International Covenant on Civil and Political Rights; General Comment 34, 2011, para. 11), may be enhanced to the extent that cross-check serves to limit enforcement against content that does not violate platform policies. This results in positive impact for the posting user and those who wish to access their content.

64. The Board also notes that cross-check could, in theory, serve to ensure that those who face particular barriers to exercise their right to freedom of expression benefit from the additional layer of protection that the program may provide. Targeted mass reporting of non-violating content, for example, could be inhibited by a false-positive mistake-prevention system.

65. However, these positive effects may be limited if the system is designed primarily to protect or prioritize the expression of people who are already powerful. The Board also notes that the cross-check program raises non-discrimination concerns, as certain entities are afforded additional protection.

66. Further, the cross-check program's protection of violating content may contribute to an environment that inhibits expression from those who may be targeted by that violating content. The range of violating content that may be left on the platform for additional time could severely impact a variety of human rights, and the consequences may vary depending on the affected users' situations. Adverse human rights impacts will likely be felt more acutely by individuals and groups who face marginalization and discrimination.

67. The Board's analysis accounts for these standards. Its policy recommendations also acknowledge the limitations of Meta's ability to moderate content at scale. If Meta's moderation more accurately moderated the content of all users, it would not need special programs based on entitled entities to help advance its respect of human rights.

Meta's values

68. International human rights standards set parameters on Meta's policies and practices. Within those standards, however, social media companies may adopt different rights-respecting approaches. Meta's values should guide the company's discretionary decisions.

69. Meta has stated that it has five values that influence the development of enforcement of its content policies on Facebook and Instagram. These values are "Voice," "Authenticity," "Privacy," "Safety," and "Dignity." According to Meta, "Voice" is the company's "paramount" value. The Board finds that cross-check, and a false-positive mistake-prevention system in general, primarily engages "Voice," "Privacy," "Safety," and "Dignity."

70. A false positive mistake prevention system that keeps content on the platform that does not violate Meta's policies contributes to Facebook and Instagram as places for expression. Conversely, to the extent that a false positive mistake-prevention system keeps violating and harmful content on the platform and facilitates its reach, it may negatively impact the "Voice," "Safety," "Privacy," and "Dignity" of others. To the extent that the system privileges the speech of some relative to others by delaying and decreasing the probability of enforcement, this unequal treatment implicates Meta's value of "Dignity," which relates to the expectation that Meta will treat all users fairly. The company should ensure that its systems are structured to consider the full range of Meta values.

V. Assessment of cross-check system

71. At the time of the Board's briefings with Meta, it performed about 100 million enforcement attempts on content every day. At this volume, even if Meta were able to make content moderation decisions with 99% accuracy, it would still make one million mistakes every day. Meta's content moderation mistakes include over-enforcement and under-enforcement, meaning that Meta both removes non-violating content and fails to remove violating content.

72. In this respect, Meta's use of cross-check responds to broader challenges in moderating immense volumes of content. The Board agrees that within this challenging context, Meta needs mechanisms to address both false positives and false negatives. Meta has a responsibility to address these larger problems in ways that benefit all users and not just a select few, however. Any decisions related to delaying or exempting enforcement actions for either certain users or certain pieces of content should align with Meta's human rights responsibilities and its stated values. Cross-check, both as it previously operated and in its current form, fails to do that.

73. The Board notes that Meta has made improvements to this system, both before referring this request to the Board and during the time the Board has been assessing cross-check. However, several aspects of the cross-check system do not

align with Meta's responsibility to identify and mitigate negative human rights impacts or uphold the company's values. These include:

- A broad scope to serve multiple and contradictory objectives that enables visibility and virality for violating content.
- The unequal access to discretionary policies and enforcement.
- That program enrollment may exceed capacity.
- The failure to track core metrics to assess the program and make improvements.
- The lack of transparency and auditability about its functioning.

74. Despite significant public concern about the program, Meta has not effectively addressed problematic components of its system. In this section, the Board highlights several of these problems. In the sections that follow we make a series of recommendations to Meta to outline how a mistake-prevention system could better comply with the company's commitments.

<u>Broad scope to serve multiple and contradictory objectives that enables visibility for violating content</u>

75. Meta told the Board that Early Response Secondary Review exists to "protect voice, [and] enhance transparency and community trust." Meta further called attention in its request to the Board to its inclusion of journalists and community leaders in cross-check. The company highlighted that cross-check ensures that voice is preserved in a variety of important scenarios:

- "Members of marginalized communities who re-share violating hate speech targeted at them in order to raise awareness about or condemn it, which have been mistakenly removed for violating our hate speech policies."
- "Journalists who report in conflict zones where designated organizations are active, whose content has been mistakenly removed for violating our Dangerous Organizations and Individuals policies."
- "Health-related nudity, such as post-mastectomy reconstruction or breastfeeding photographs, which have been mistakenly removed for violating our nudity policies."

76. In a meeting with the Board, when asked about negative impacts that might ensue without ERSR, Meta officials stated that one issue, for example, is that it could impede communication and the flow of information in a crisis such as a natural disaster or political upheaval. These points of emphasis in Meta's stated rationale for the system contrast strikingly with how the system operates.

77. The Board shares Meta's concern about wrongly removing non-violating content posted by people drawing attention to human rights violations, working to promote women's health, and other public interest reporting. In fact, the Board's decisions have addressed such mistakes. Meta identifies these cases as "enforcement errors" only after the Board brings those cases to the company's attention. Examples

include the *Wampum belt* decision (2021-012-FB-UA), concerning the incorrect removal of an Indigenous artist's expression countering hate after multiple erroneous human review decisions; the Board's *Mention of the Taliban in news reporting* decision (2022-005-FB-UA) about the incorrect removal of a news outlet's post reporting on a designated organization; and the *Breast cancer symptoms and nudity* decision (2020-004-IG-UA), concerning the incorrect automated removal of a post that should have benefited from the health-related exception to Meta's adult nudity policies.

78. While Meta focuses on at-risk voices posting non-violating content when describing the program, Meta also stated that the cross-check program serves a core business function, as it serves an "important role in managing Facebook's relationships with many of [its] business partners." Relatedly, the cross-check tag sensitivity framework, which underpins both the "entity sensitivity" factor for GSR ranking and ERSR tags, is directly linked, among other factors, to the degree of reputational and internal backlash that is anticipated if particular content is removed in error. For example, Meta characterizes the risk of "escalation at the highest levels (CEO, COO)" as corresponding to a cross-check "extremely high severity" tag. Correlating highest priority within cross-check to concerns about managing business relationships suggests that the consequences that Meta wishes to avoid are primarily business-related and not human rights-related.

79. In order to assess how Meta prioritizes entities within cross-check, the Board repeatedly requested that Meta share its Early Response Secondary Review list for the Board's analysis. Meta did not provide the Board with this list. The Board cannot fully assess the degree to which the company is meeting its human rights responsibilities under the program or the profile of the entities that are guaranteed enhanced review if it does not know how the program is being implemented and precisely who benefits from it. Meta argued that providing a list of users subject to cross-check would violate the company's legal obligations in relation to user privacy. Based on legal advice, the Board believes, and has pointed out to Meta, that these concerns could have been mitigated and more extensive disclosures provided.

80. Almost five months after the Board first requested this information, Meta provided the Board a list with limited aggregate data about each listed entity on the current Early Response Secondary Review list. Specifically, Meta only disclosed the type of entity (e.g., Instagram user, Facebook page), their associated country and language, as self-selected by the entity, and whether or not Meta considers the entity "civic" and a "partner." Not all information was provided for each category of entity. For example, a quarter of the listed Instagram entities did not select a specific country or language in their profile settings and are also not considered either a civic actor or Meta partner.[50] This means that, for these entities, Meta only disclosed the existence, but not the identities or characteristics, of a group of Instagram users benefitting from cross-check.

81. This limited disclosure impairs the Board's ability to carry out its mandated oversight responsibilities. Meta's description of the "civic" category, for example, includes state actors, elected officials, "civic influencers" and candidates to public office, among others. Similarly, the "partner" category spans news organizations, celebrities, artists and more. The Board cannot evaluate, for example, the degree to which journalists, rights defenders and dissidents in particular countries are granted the same protection for their expression as the state actors who are enrolled in ERSR under program policy.

82. Meta has told the Board that it has no comprehensive system in place to systematically assess which journalists, human rights defenders or civil society figures in a particular geography should be subject to ERSR. Inclusion of such users on the list is based on decentralized decisions by Meta staff described by the company as "internal experts with high market knowledge." This raises the risk that significant gaps and inconsistencies exist in terms of who is afforded the added layers of protection for expression that cross-check ERSR provides.

83. Journalists posting content from conflict contexts, political opposition seeking elected office, celebrities posting a wide range of content, and business partners posting content to sell goods pose fundamentally different risk profiles from a free expression and human rights perspective. Given Meta's problems moderating content at scale, within current limitations user-generated content should be subject to different rights-oriented prioritization. Meta has described a system that does not include strategies or tactics to ensure that the individuals and the expression most needing protection receive it in the near term, with the ultimate goal of providing better content moderation for all.

84. Under ERSR, should content from any entitled entity be identified as violating and flagged for additional review, such content, regardless of risk profile, remains on platform during its period of peak virality in the aftermath of immediate posting. This is significant for two reasons. First, viral content spreads quickly on and across platforms. Second, once something is posted by an entity that has a large reach, the content will inevitably be recorded and reshared individually by users even if the original post is deleted. This means that accounts that benefit from the ERSR cross-check system may upload violating content and know that it can attain far reach even if it is violating.

85. Although the Board notes that Meta stated it has a system to prioritize high severity ERSR content for review, this content still remains on the platform until all necessary reviews are completed, sometimes for significant periods. For example, in the Neymar case, it is difficult to understand how non-consensual intimate imagery posted on an account with more than 100 million followers would not have risen to the front of the queue for rapid, high-level review if any system of prioritization had been in place. Given the serious nature of the policy violation and the impact on the victim, this case highlights the need for Meta to adopt different approaches for content pending review and shorten review timelines.

86. Delayed enforcement of violating content is a significant source of harm under the cross-check program. According to Meta's own research, user views of violating content because of cross-check are due to "incorrect overturns, and the delay of enforcement of non-overturns for which enforcement is slowed due to the secondary review process." The company acknowledges that affording additional protection to some privileged users' content may confront other users with violating content such as hateful speech or harassing posts.

87. Content automatically granted ERSR is different from content identified and sent for GSR. On the one hand, as noted above, the percentage of ERSR content ultimately found violating seems to vary. During time periods when the overturn rate is low, a key flaw in the system is its failure to ensure the prompt removal of violating content.

88. On the other hand, most GSR content is consistently ultimately found non-violating. For this system, the overturn rate seems to reveal that there are greater over-enforcement issues at scale, and that secondary review is mostly permitting non-violating content to remain accessible. The Board thus notes that to the extent that GSR preserves more expression, its impact is limited by the capacity constraints Meta imposes.

89. In sum, the Board finds that while Meta characterizes cross-check as a program to protect vulnerable and important voices, it appears to be more directly structured and calibrated to satisfy business concerns. While the Board understands that Meta is a business and should be able to design policies that meet business concerns, these same policies should not be characterized as serving as human rights risk mitigation measures if they do not meet that objective. Additionally, if Meta's business design choices negatively impact human rights, it should identify and then prevent, mitigate, or cease those negative impacts through program improvements.

<u>Unequal access to discretionary policies and enforcement</u>

90. Cross-check is designed to subject some content to more nuanced moderation decisions, determining whether any exception or specialized policy might apply to decline enforcement. According to Meta, "if content that was cross-checked is escalated for further review, it may then be subject to a decision based on [...] context-specific policies." Cross-check enables human review by the "Early Response Team" which the Board believes can grant exceptions in enforcement, both relating to the specific content and penalties against the entity itself. Content reviewed through ERSR is guaranteed to reach this team before possible removal, and content reviewed through GSR has a higher chance of reaching this team.

91. Meta has repeatedly told the Board and the public that the same set of policies apply to all users. Such statements and the public-facing content policies are

misleading, as only a small subset of content reaches a reviewer empowered to apply the full set of policies.

92. Entitlement to Early Response Secondary Review therefore provides a significant benefit to the user. It means that more of the content they choose to post is more likely to remain on the platform. In the case of non-violating content, it is protected from mistaken removal. In the case of violating content, it is allowed to remain on the platform during peak viewership before a later removal.

93. The Board also believes that, in addition to applying content policies with more discretion, content reviewed on escalation may benefit from decisions to not apply account restrictions that would be applied under normal procedures. In general, content policy violations correspond to "strikes" against an account, which in turn correspond to specific consequences. According to Meta's Transparency Center, strikes lead to increasingly long periods of time where accounts cannot post content. For serious or repeated strikes, Meta will disable an account.

94. The Board inquired about the discretionary application of policies and enforcement consequences. Meta responded that it does "not have statistically significant data distinguishing between penalties applied to cross-check versus non-cross-checked entities" and is "not aware of and [has] not located research or analysis" addressing these possible discrepancies. Given that cross-check may exempt users from account-level consequences, the Board is troubled that the company has either chosen not to track and analyze this information or has failed to disclose it to the Board.

95. According to public reporting in The Guardian, after Neymar posted violating content, he "was not subject to the normal Facebook procedure for someone who posts unauthorized nude photos, which is to have their account deleted." This example was revealed through whistleblower disclosures, and it is not clear how widespread such practices may be. The Board also asked Meta to confirm the account-level restrictions it applied in this case. The company ultimately disclosed that the only consequence was content removal, and that the normal penalty would have been account disabling. The Board notes that Meta later announced it signed an economic deal with Neymar for him to "stream games exclusively on Facebook Gaming and share video content to his more than 166 million Instagram fans."

96. Unequal access to escalated reviews as well as policy exceptions is particularly concerning given the lack of objective or transparent criteria for inclusion on Early Response Secondary Review lists. As noted above, it is not clear how Meta ensures that those most likely to be subject to over-enforcement or facing challenges to exercising their rights to freedom of expression are given this additional protection. The Board is concerned that those often most at risk, including journalists and human rights defenders, who may report on dangerous organizations or document graphic abuses, are those least likely to be proactively added to such lists given the investment that would be required to find those people across the globe.

97. On the other hand, Meta explained to the Board that it has a dedicated team charged with ensuring that all eligible entities representing government officials and organizations are enrolled in ERSR. Criteria to include "businesses, media organizations and creators" also seem clearer. According to Meta, one criterion, for example, is a specific amount of spending or revenue generated by an entity across Meta's "family of apps," although the amount may vary over time.

98. The Board is also concerned that in operating cross-check, Meta focuses disproportionate attention on more lucrative markets, instead of focusing on contexts with greater risks to human rights, including freedom of expression. At the time of the Board's briefing with Meta, 42% of content reviewed through the Early Response Secondary Review pathway originated from the United States or Canada. Similarly, 20% of all the entities on ERSR lists at that time correspond to those two countries. In contrast, according to Meta, just 9% of "monthly active people" on Facebook were from the United States and Canada. This data shows that users based in the United States and Canada have disproportionate access through Early Response Secondary Review to specialized review pathways that guarantee access to the full set of Meta policies, analysis of context, and likely the possibility of non-standard account penalties for violating content.

99. This disparity is correlated to the fact that "average revenue per person" in the US and Canada is the highest in the world, at around three times larger than in Europe and about 12 times larger than in Asia-Pacific. These facts highlight the financial incentives that shape how ERSR operates and reinforce concerns of equity. Through the design of cross-check, users in lucrative markets with heightened risk of public relations implications for Meta enjoy greater entitlement to protection for their content and expression than those elsewhere.

100. In addition, for GSR, the cross-check ranker prioritizes content according to factors like "topic sensitivity" that potentially require automated assessment of the language of the content. The Board is concerned that Meta does not prioritize training its automated processes on less-spoken languages and less lucrative markets. Limited investment in moderation in these languages limits the ability of algorithms to identify topics in such content. This suggests that users in these markets, including the Global South, may be disadvantaged when assessed for GSR cross-check eligibility. Similarly, Meta disclosed that "a group of languages are reviewed by non-native speakers using our translation and slurs highlighting tools." This reinforces the Board's concern that cross-check does not benefit all users equally, even through GSR.

Program enrollment exceeds capacity

101. ERSR and GSR eligibility persistently exceeds the human review capacity Meta allocates to the cross-check program. The mismatch between the volume of content that is designated for enhanced review through these systems and the

inadequate human resources allocated to the task represents a critical flaw in the system.

102. Meta told the Board that it "never intended to operate with a consistent backlog of cases, though operational capacity constraints and increasing volumes have led to a backlog in Early Response Secondary Review. [...T]hat backlog consists of content we have assessed as likely being low severity." Notwithstanding Meta's statement that it did not intend to maintain a continuous backlog, the company has failed to assign sufficient human resources to meet the content moderation needs of these programs. Moreover, as noted above, not all content subject to delayed enforcement is low severity.

103. Limited human review capacity has different but related consequences for Early Response Secondary Review and General Secondary Review. For ERSR, capacity shortfalls mean that content will remain on the platform during the time period it is most likely to accrue views. As this content remains on the platform until it receives enhanced review, content posted by high-profile ERSR users that violates Meta policies remains on the platform during its period of highest viewership. While Meta may attempt to review content that may cause greater harm first, it is not clear that it does so consistently, and this still reflects a design decision to provide automatic protection to entities selected based largely on commercial criteria.

104. For General Secondary Review, limited capacity may lead to two consequences. First, the Early Response Team may fill its time with ERSR content, as that content must be reviewed to apply any enforcement action. The Early Response Team therefore often does not have availability to review GSR content, and GSR content does not reach this critical level of review where policies requiring additional context and discretion may be applied. Second, limited capacity at the Market Team review level means that more GSR content times out of the queue before review and is removed by default. As the majority of this content appears to consistently be non-violating, this means that a key consequence of limited capacity for General Secondary Review content is that Meta removes more content that is likely non-violating.

105. These flaws compound the disparities in treatment of different users on the platform. Privileged users enrolled in ERSR have more chances to be reviewed by a moderator who may apply context to uphold their content, have a greater range of policy exceptions that can apply to uphold their content, and benefit from a system where even violating content is guaranteed viewership for some period of time. Ordinary users whose content might have access to GSR review, by contrast, have more limited opportunities for review of their content, are more likely to be subject to content policies without contextual review or the applications of policy exceptions, and, as content times out, are more likely to have non-violating content removed. This system has serious implications for the values of "Voice," "Dignity," "Privacy," and "Safety" that Meta claims to pursue.

Failure to track core metrics to assess the program and make improvements

106. The Board evaluated the metrics Meta uses to justify and evaluate the cross-check program. The metrics that Meta currently uses do not capture all key concerns and do not seem to have provoked changes when shortcomings were identified. Additionally, Meta is failing to monitor and set goals on a broad enough set of metrics to give a full picture of how the program operates and establish targets for improvement accordingly.

107. As discussed above, one metric Meta calculates is the overturn rate, or percentage of content that is subject to cross-check and ultimately found non-violating, despite the initial identification as violating by automation or human review. In Meta's words, "overturn rate is the efficacy rate of the cross-check system." According to information it provided the Board, Meta "want[s] [the overturn] percentage to be high. If none of the decisions were overturned through cross-check, that would mean [it was] cross-checking the wrong content."

108. Even though Meta has stated that the overturn rate should be high, Meta continues to provide the greatest protections to Early Response Secondary Review users. According to the figures Meta provided the Board, this rate varies significantly. Providing this protection to content without a consistently high overturn rate suggests Meta may be, according to its own goals, cross-checking the wrong content.

109. Content moderation in large volumes is marked by both over and under-enforcement. Meta focuses on the prevalence of violating content as its main public metric to assess how effective its moderation efforts are at removing harmful content. This includes content that flows through the cross-check system. Meta calculates prevalence by estimating the percentage of all views of content on Facebook or Instagram that were views of violating content. The use of prevalence as its general success metric may encourage Meta to further automate the removal of content and limit context-based enforcement to ensure low prevalence across the platform, without proper mechanisms to prevent wrongful deletions of content at scale. The consistently high overturn rate on GSR, for example, supports that inference.

110. The Board notes that Meta did not provide the Board with information showing that it tracks data about the accuracy of decisions made through its cross-check system. This means that even though the program is supposed to ensure accurate content moderation decisions, it does not appear that Meta is tracking whether its decisions via cross-check pathways are more or less accurate than decisions made through its normal scaled quality control mechanisms. Accuracy data would be a key indicator of the possible influence of non-content policy concerns on moderation decisions made in cross-check. By measuring success based on overturn rate only, Meta is not considering whether the ultimate decisions are the correct decisions.

111. Additionally, Meta has stated that the Regional Market Teams and the Early Response Team are specialized, having a particular set of skills, training and access to internal tools that allows them to make cross-check-level moderation decisions. However, as described above, at certain points in both the ERSR and GSR pathways, decisions may be made by contracted reviewers. These reviewers do not have the same access or training as Meta employees. If the goal of the cross-check program is to produce the most policy accurate decisions for entitled entities and important content, then measuring accuracy of cross-check decisions in general, but across reviewer types in particular, should be a basic tenet to understand if the operational design is working as intended.

112. Additionally, as an objective of cross-check is to protect important content most at risk of over-enforcement, Meta should focus on additional methods to identify such content. Meta disclosed that while it is actively working on understanding and mitigating over and under-enforcement for specific populations and problem areas, it still "needs to centrally define which populations are under/over-enforced. Pending such an effort we do not have a good way of creating a before the fact definition."

Lack of transparency and auditability of the program and its functioning

113. Lastly, the Board is concerned about the limited information Meta has provided the public and its users about this program. This policy advisory opinion resulted from Meta's failure to disclose to the Board key information about this program in the context of its deliberation on a case about a prominent user subject to cross-check.

114. Currently, Meta does not inform users that they are subject to ERSR, the entity-based mechanism in cross-check. It also does not inform users when they report content posted by a cross-checked entity. The company also provides limited transparency on the complex secondary review processes cross-checked content benefits from.

115. In addition, Meta does not share publicly its procedures for ERSR list creation and its auditing framework. The Board does not know, for example, whether entities that continuously post violating content are kept on Early Response Secondary Review lists based on their profile. Meta has given no indication that violation history or frequency is a factor in creating or maintaining Early Response Secondary Review lists. The lack of transparency regarding auditing impedes the Board and the public from understanding the full consequences of the cross-check system.

Conclusions about cross check

116. The Board acknowledges that a mistake-prevention system could be a useful safeguard against the improper removal of important content. However, if cross-check does not target such expression and permits severely violating content to remain on the platform, the program creates negative human rights impacts that

Meta is not monitoring or mitigating. The Board thus concludes that cross-check is currently neither designed nor implemented in a manner that meets Meta's human rights responsibilities and company values.

117. In its case decisions, the Board looks to the three-part test in Article 19 of the ICCPR, evaluating whether restrictions on expression meet requirements of legality, legitimate aim, and necessity and proportionality.

118. Legality refers to whether the rules are clearly and accessibly communicated. The existence, purpose and nature of the system is opaque in ways that cannot be justified given the significant effects cross-check has on the exercise of fundamental rights. Content policies presented as globally applicable that can only be applied with additional context at escalation, including through cross-check, are misleading.

119. Legitimate aim refers to whether restrictions are targeted at objectives specified in Article 19, including to respect the rights of others and protect national security, public order, and public health. The metrics through which the company measures the effectiveness of its enforcement systems suggest that its motivations are substantially focused on business reasons.

120. Necessity and proportionality refer to whether restrictions on expression are the least intrusive way to meet the legitimate aim. Here, the Board reiterates its concerns about inequitable access to the benefits of cross-check. Meta maintains clear processes to determine some of its users are entitled entities, such as state actors and business partners. Without clear criteria for other users who are likely to post content with significant human rights value, the program less clearly benefits others, including members of marginalized and discriminated-against groups. Meta is also not collecting and monitoring information about whether this program produces more accurate results in practice. Lastly, through cross-check, Meta defaults to leave content identified as violating on its platforms. As a policy matter, Meta is setting aside what it has determined is a proportionate response at scale for some content, often based only on economic or public relations concerns.

121. To comply with Meta's human rights responsibilities and company values, a system to prevent over-enforcement should be structured substantially differently than it is at present.

VI. Enforcement recommendations

122. In response to the questions posed by Meta, here the Board provides recommendations on entity-based mistake-prevention systems and dynamically determined content-based mistake-prevention systems. Meta has a responsibility to address its content moderation challenges in ways that benefit all users and not just a select few. However, given the focus of this policy advisory opinion, the Board

focuses here on limited-scope mistake-prevention systems.

Entity-based mistake-prevention system governance recommendations

123. Any system based on entity eligibility, such as Early Response Secondary Review, should be carefully designed, subject to oversight, and continuously monitored. This should ensure that it meets its stated purposes and evaluates externalities and unintended consequences it may cause. Such a system should protect users who are likely to post expression that is particularly important from a human rights perspective.

124. It is critical that Meta be clear about its objectives and tailor its systems narrowly to meet those objectives. It should also avoid providing protection for expression that violates its content policies or human rights commitments. Additionally, given that certain users may benefit from additional protections and avenues for expression, the company should provide the public with robust information about these processes so that they can adequately evaluate the information and opinions they see on the platform.

Users that should be included in entity-based mistake prevention systems

125. Meta states that the categories of inclusion for its entity-based mistake-prevention system cover "civic and government," "significant world events," "media," "historically over enforced," and "marginalized communities," "businesses," "creators," "entities escalated for review," and "legal and regulatory."

126. These broad categories require additional sorting and specification. In light of Meta's human rights commitments and stated values, if the company opts to operate an entity-based false-positive prevention system, there are certain categories of users that *should* be provided such protection, users that *may* be provided this protection, and users that *should not* be provided such protections given the human rights risks they pose.

127. First, entities that *should* be included are those who are likely to produce expression that it is important from a human rights perspective, including on matters of public importance. This benefits not only those users, but those who wish to access the information they share.

128. These users should include, for example, people whose content runs a high risk of over-enforcement, journalists and media organizations, public officials and candidates for office, and other civic actors including human right defenders and advocates for marginalized communities. In this respect, the Board views a list-based system as a proxy for providing additional protections to critical expression, and not protection based simply on the identity of the speaker. The Board recognizes that Meta maintains various lists of entities to which it affords greater protection, including its journalists' registry and its roster of "trusted partners" from

civil society. These existing, vetted entities could form one source from which the company might build an objective, global, human rights standards-based system accessible to all those whose expression satisfies the criteria for inclusion.

129. Second, entities that *may be* included may be based on company priorities and may include users with commercial importance and business partners. This might include advertisers, businesses with pages or groups that are at risk of over-enforcement, users who pose a special reputational risk to the company, or other users with a commercial relationship with Meta.

130. Third, there are entities that *should not* be included in any entity-based mistake prevention system that delays all enforcement. These include entities and users that repeatedly create or share content that violates Meta policies or terms of service. Meta's current account-level enforcement system, based on strikes and penalties, could be leveraged for the purposes of implementing this rule. Should users included due to commercial importance frequently post violating content, they should not continue to benefit from a system that delays enforcement. Meta has a responsibility to identify such users and exclude them from systems that provide their violating content additional visibility. While the number of followers could be a legitimate proxy for the degree of public interest in user's expression, a user's celebrity or follower count should not be the sole criterion for an entity-based mistake prevention system.

131. Meta's inclusion of all entities in the same system places them in direct competition for limited review resources. Meta should prioritize adequately resourcing mistake-prevention systems that mitigate human rights harms. In this context, Meta should ensure that content with human rights or public interest implications is reviewed in a timely fashion by skilled reviewers with the ability to take further context into consideration, regardless of whether the content came from entity-based or content-based pathways.

132. The Board recommends that Meta take steps to either use separate pathways or create prioritization mechanisms to differentiate between users that should be included due to Meta's human rights responsibilities and users that are included due to commercial priorities, given their different risk profiles. Businesses, for example, might be more likely to have content identified as violating spam rules, as they may rapidly post commercial content. Users with a large follower count may post on important matters of public interest but may similarly post violating content.

Decision makers should be qualified and empowered to make rights-respecting decisions

133. Consistent with the Board's recommendations throughout, Meta should prioritize its mistake-prevention secondary review workflows according to risk profile and human rights value.

134. The content posted by entities that Meta *should* include based on human rights concerns should be reviewed by teams with context and language expertise. This review pathway, including its escalation paths, should be devoid of business considerations. Meta should take steps to ensure this team does not report to public policy or government relations teams or those in charge of relationship management with any affected users.

135. The path dedicated to resolving issues explicitly related to Meta's business priorities could address, for instance, ads enforcement, spam rules, feature limits and behavioral issues. An example of behavioral issues is a business page being wrongly penalized for uploading pictures at a much faster rate than a normal profile. Either through lower prioritization or separation into a different workflow, these reviews should not displace resources targeted at human rights mitigation.

136. The Board notes that the Early Response Team, which is permitted to apply policy exceptions and interpret context, does not require its reviewers to have cultural or linguistic expertise. According to Meta, it makes decisions based on notes provided by Regional Markets Teams. Meta itself acknowledged that "relying on translations is imperfect." In this context, the Board urges Meta to ensure cultural and linguistic expertise at these levels of review. Meta should consider incorporating employees with cultural and linguistic expertise from at-risk regions into these teams and developing procedures to include staff with such expertise in decision making.

Guidance to create and govern lists for entity-based mistake-prevention systems

137. Meta should establish clear and public criteria for entity-based mistake-prevention eligibility. These criteria should differentiate between users whose expression merits additional protection from a human rights perspective, including information in the public interest, and users included for business reasons. For example, Meta currently defines one cross-check category as "Media Organizations, Businesses, Communities and Creators." This category includes "health organizations, news publishers, entertainers, musicians, artists, creators, and charitable organizations." Criteria this broad are insufficient. Meta should also develop criteria based on patterns of violating or undesirable behavior on the platform to avoid granting protections to harmful users.

138. Meta should add entities to mistake-prevention systems only once the process is objective, well-governed, and transparent. All entities that are proposed to be added to a list should be made aware of the possibility and should be given the option to decline inclusion if they so desire. Those who choose to be included should be required to review Meta's content rules and re-commit themselves to following them. While the Board views cross-check as providing benefits to included users, Meta should operate based on principles of user consent.

139. Clear public criteria should also provide a basis for users who qualify to proactively seek inclusion on such lists. Meta should establish a process whereby users can

apply for over-enforcement mistake-prevention protections should they meet the company's articulated criteria. State actors should be eligible to be added or apply based on these criteria and terms but given no other preference.

140. In addition to meeting public criteria, the process for inclusion, regardless of whether a user or Meta initiates the process, should involve: (1) a requirement to review Meta's content policy and an additional, explicit, commitment to follow them; (2) an acknowledgement of the program's particular rules; and (3) a system to inform users proactively of changes to Meta's content policies to facilitate awareness and compliance.

141. Meta sometimes works with civil society through its 'trusted partner' program and other stakeholder engagement initiatives to gather information about entities that should be considered for protection. The Board recommends that Meta strengthen its engagement with civil society for the purposes of list creation. Users should be able to nominate others that meet the public criteria, as long as the nominees may decline inclusion. This is particularly urgent in countries where the company's limited presence does not allow it to identify candidates for inclusion independently.

142. List creation, and particularly this engagement, should be run by specialized teams, independent from teams whose mandates may pose conflicts of interest, such as Meta's public policy teams. To ensure criteria are being met, specialized staff, with the benefit of local input, should ensure objective application of inclusion criteria. Public policy teams often interact with and lobby government actors, creating unavoidable conflicting incentives. While they may nominate candidates, they should not be decision makers.

143. Meta told the Board that currently a single company employee may decide to add entities to a particular cross-check list, and there is no required review of those decisions. Going forward, the company should have an established process for objective, criteria-based review of all entities that will receive additional benefits. At least two people on different teams should be involved to finalize inclusion on any list-based protection, and individuals with personal or business relationships with nominated entities should not be decision makers.

Guidance to maintain and audit lists for entity-based mistake prevention systems

144. In addition to establishing clear criteria for entry to a mistake-prevention protection program, Meta should establish clear criteria and processes for audit and removal. Should entities no longer meet eligibility criteria, they should be removed.

145. Meta told the Board that its new proposed governance structure includes rules to add and remove entities from the lists; tag expiration rules; periodic audit procedures; and an oversight structure. Meta also disclosed, however, that there were exceptions to some of these rules, such as "Civic and Government" entities not

having default expiration periods. Meta also shared that it is currently auditing a limited subset of entities on Early Response Secondary Review as it moves towards a more simplified list structure.

146. The Board recommends that Meta require at least yearly review of all included entities in any mistake-prevention system that provides benefits to such entities. There should also be clear protocols to shorten that period where warranted. Similar to its recommendations on initial inclusion in any list-based system, the Board recommends that at least two people with separate reporting structures participate in internal audits.

147. Meta should also ensure clear removal criteria for any list-based protection program. One criterion should be the amount of violating content posted by the entity. These could, for example, be based on a "three-strikes" policy, unless Meta has established a harsher penalty for the violation(s) at issue (e.g., Non-Consensual Intimate Imagery account removal). Such a system should give entities warnings and then remove them from cross-check when they accrue their final strike, regardless of whether the violation merits removal from the platform as a whole. Entities should be able to appeal removal and reapply in the future.

148. Lastly, the Board emphasizes that, while internal audit procedures are a step in the right direction, internal auditing without external oversight falls short. External audits, by the Board or another third party (e.g., researchers or civil society), are required in order to assess whether a mistake-prevention system mitigates negative human rights impacts. While the Board acknowledges serious privacy and safety concerns with external auditing, the Board believes that Meta can take mitigating steps to anonymize and aggregate data to address these concerns.

Some entities receiving additional protection should be publicly marked

149. The Board has repeatedly called on Meta to inform users and the public about its policies and practices. Any entity-based mistake-prevention system should provide all users on the platform with clarity about how Meta applies its rules. Currently, users do not know if they are enrolled in ERSR. Additionally, users viewing and reporting content posted by users enrolled in ERSR are not informed that the content may be subject to special review procedures.

150. The Board recommends that some categories of entities protected by the system should have their accounts publicly marked. These categories include all state actors and political candidates, all business partners, all media entities, and all other public figures included because of the commercial benefit to the company in avoiding false positives. This will allow the public to hold privileged users accountable for whether protected entities are upholding their commitment to follow the rules and hold Meta accountable for adhering to the publicly announced parameters of the program.

151. The Board identifies several risks in publicly identifying users who are enrolled in a false positive mistake-prevention program. First, there could be additional adversarial risk from users attempting to gain control of accounts with special protections, knowing that violating content will remain on the platform for a period of time. Second, some categories of users may face harassment or other attacks if they are perceived as maintaining a relationship or receiving special protection from the company.

152. However, the Board finds these risks can be mitigated, and the benefits outweigh these potential harms. First, Meta should invest any necessary resources to enhance account protection for users subject to a mistake-prevention system. Meta has experience providing extra layers of protection for journalists and other categories of users. Such procedures could be adapted for use in any future entity-based mistake-prevention system. While adversarial risk is real, it is not insurmountable in this context. Although the Early Response Secondary Review list is currently not public, many users already assume that high-profile accounts are included in the cross-check program.

153. Meta should not identify beneficiaries who are human rights defenders, entities included because they are subject to historical over-enforcement, and those included because they are at risk of harm, although they should be able to opt-in to identification. Clear criteria for inclusion and separation of the program for different objectives will facilitate this process.

154. Lastly, when users report content posted by an entity publicly identified as benefiting from additional review, reporting language should make it explicit that special procedures will apply, explaining the steps and potentially longer time for resolution.

<u>Content-based mistake-prevention system governance recommendations</u>

155. While an entity-based system should include users who are likely to produce expression that merits additional protection from a human rights perspective and users who may be at particularly high-risk for erroneous over-enforcement, a content-based system seeks to protect such content directly without regard to who posted it.

Content that should be selected and prioritized for content-based mistake prevention systems

156. According to Meta, its General Response System "ranks content based on false positive risk using criteria such as topic sensitivity (how trending/sensitive the topic is), enforcement severity (the severity of the potential enforcement action), false positive probability, predicted reach, and entity sensitivity (based largely on the compiled lists, described above)."

157. The most heavily weighted factors for the ranking algorithm are topic sensitivity and entity sensitivity. As discussed above, entity sensitivity is, among other factors, directly related to the degree of internal escalation a mistake would cause. In this respect, Meta's cross-check ranker also prioritizes content that might cause economic or reputational damage, an objective already served by ERSR. Although GSR may have been designed to respond to some criticisms of the former exclusively entity-based cross-check system, this suggests that the company continues to prioritize expression based on the speaker and not the importance of the expression.

158. The Board agrees that universal eligibility for a false-positive mistake-prevention system is a positive step. However, such a system should prioritize identifying content that is not also targeted by an entity-based system. It should provide enhanced protection based upon a human rights rationale. While Meta may give some additional protection where over-enforcement might threaten its business interests, similarly to list-based systems, it should not do so at the expense of its human rights commitments.

159. An algorithmic ranker for a false positive prevention system could, for example, prioritize content based on the types of decisions that are hard for automation and human moderators at scale (e.g., historically over-enforced speech or speech by marginalized communities). In conjunction, the algorithm could prioritize the review order of this content based on the severity of the possible violation, the likelihood of being a false positive, and the likelihood of virality.

160. The Board recommends that to increase the impact of a content-based false positive prevention system, Meta should consider reserving a minimum amount of review capacity by teams that can apply all content policies (e.g., the Early Response Team). It further should analyze the content receiving additional review for insights as to where Meta's systems are resulting in the most high-impact mistakes, and prioritize review resources accordingly.

Technical corrections

161. Meta explained that "technical corrections" completely prohibit any enforcement for a specific policy violation on a particular entity. There may be business reasons to provide such protection to an extremely select set of entities, but any such system has the potential to create great risk of exempting entities that post violating content from content moderation enforcement. If such a system is used, it should be subject to the highest level of internal and external scrutiny. "Technical corrections" exempt certain entities from certain enforcement and are rightly understood as an "allowlist" or "whitelist," however limited its scope may be.

162. All recommendations regarding list-based programs, such as clear and firm criteria, cross-team review processes to grant any exemption, and audit processes to maintain exemptions apply here. Additionally, exemption should be prohibited for

content that Meta classifies as a high-severity violation. Meta should conduct periodic audits for all enforcement actions that are blocked by such exemptions. If, as Meta states, this is about a thousand actions per day, it should have the capacity to do so. This audit, with information on the scope and accuracy of the program, should be included in Meta's quarterly transparency reports.

163. Finally, the company should proactively and periodically search for unexpected or unintentional exemptions that may linger from previous iterations of this program. In its decisions, the Board repeatedly notes cases where Meta has inadvertently failed to update or maintain systems, and the consequences of such gaps in governance on an exemption system could be critical.

General mistake-prevention system governance recommendations

164. Beyond the broad governance changes to how a list-based and content-based mistake-prevention system should be established and audited, the Board also recommends that the procedures within such a system focus on harm mitigation and be subject to continuous monitoring for learning and improvement.

Harm mitigation following identification of violating content

165. As the company itself recognizes, a core cause of harm in Meta's false-positive mistake-prevention system results from delayed enforcement of violating content during the time period in which it is most likely to be viewed. As stated above, Meta itself identifies that the biggest drivers of users seeing violating content from cross-checked users or content on its platforms are "incorrect overturns, and the delay of enforcement of non-overturns for which enforcement is slowed due to the secondary review process." The Board urges Meta to take steps to mitigate those harms.

166. First, Meta should take measures to ensure that additional review decisions are taken as quickly as possible. Investments and structural changes should be made to expand the review teams so that reviewers are available and working in relevant time zones whenever content is flagged for any enhanced human review.

167. Second, the Board recommends that Meta use methods aside from defaulting to no enforcement action for pieces of content subject to enhanced review. This could include using the least intrusive means, for example, downranking, slowing the virality, hiding, or temporarily removing the content. Establishing different prioritization or pathways for content and entities of different nature should facilitate Meta applying different consequences to different types of content.

168. Content identified as violating during Meta's first assessment that is high severity, for example according to Meta's framework, should be removed or hidden pending review, and not permitted to remain on the platform accruing views simply because the posting user is a business partner or celebrity. The difference between

enforcement options, such as removal, hiding, and downranking, should be based on violation severity. Meta's framework, in theory, is designed to account for the likelihood of near-term harm, and whether the content has been identified as particularly likely to be an enforcement error. If content is hidden on these grounds, a notice indicating that it is pending review should be provided to users in its place.

169. Third, Meta should not operate these programs at a backlog. Maintaining a content queue for review that exceeds capacity means that content that may be violating will remain on the platform for an extended period of time. Delaying any enforcement while taking weeks to reach a decision results in functionally exempting entitled entities from the rules.

170. Meta should invest the resources necessary to match its review capacity to the content it identifies as requiring additional layers of review. This does not, however, mean that it should have the algorithm select less content. The consequence of Meta failing to build sufficient review capacity should not be delaying content from enforcement or outright mistaken deletions made by at-scale systems or reviewers. Meta has devised processes to prioritize review and ensure that its workforce has a continual stream of content to review. Given that GSR review currently results in a consistently high overturn rate, the Board believes that more content would benefit from this review.

171. Fourth, Meta should not automatically prioritize entity-based secondary review and make a large portion of the algorithmically selected content-based review dependent on extra review capacity.

Ensuring appeal availability

172. Meta informed the Board that it does not provide appeal or review opportunities consistently across all types of content. Appeals for content subject to the cross-check program seem to suffer from the same inconsistency.

173. The Board understands that providing appeals on content that has already reached the highest level of analysis within the company may be unnecessary, given that an appeal would replicate those pathways. However, the Board is concerned that some content may not be receiving appeal eligibility, despite not reaching those highest levels. The Board believes that Meta could have and should have provided more clarity on this point when asked repeatedly by the Board.

174. Additionally, the Board is particularly concerned about this confusion as it relates to appeals eligibility to bring cases to the Board, both for users to restore their own cross-checked content and to report the content of other users that benefit from cross-check. In fact, according to Meta, "for the months of May and June 2022, an average of 35% of the content in the cross-check system [...] could not be escalated to the Oversight Board." Users included on Early Response Secondary Review lists are among the users with the highest reach on the platform. This situation may be

depriving the Board of some of the most critical content moderation cases on Facebook and Instagram.

175. As a first step, Meta must provide clarity regarding appeals eligibility in general, and ensure that content that does not reach the highest level of review is able to be appealed internally. Second, Meta must guarantee that it is providing an opportunity to appeal to the Board for all content the Board is empowered to review under its governing documents, regardless of whether the content reached the highest levels of review within Meta.

Learning and improvement

176. To meet its human rights responsibilities, Meta should monitor its activities which impact rights on a periodic basis. The results of these reviews should guide Meta in making improvements to its policies and practices to minimize human rights harms. In this case, Meta maintains a variety of metrics related to the cross-check program that already show where the company should be making improvements. The Board believes that Meta should also provide the public with information about how this system is functioning, both to meet transparency responsibilities and to hold itself accountable for improvement.

177. First, Meta already maintains an overturn rate for its entity-based system (Early Response Secondary Review) and for its content-based system (General Secondary Review). Meta should use trends in overturn rates to inform whether to default to the original enforcement within a shorter time frame or what other enforcement action to apply pending review. If overturn rates are consistently low for particular subsets of policy violations or content in particular languages, for example, Meta should continually calibrate how quickly and how intrusive an enforcement measure it should apply.

178. Second, Meta told the Board that it has conducted post-mortem analysis exercises after Meta's "risk assessment" team has identified risk areas or there was an event that the company saw as a failure. The Board recommends that these and other reviews be conducted regularly on cross-check, based on internal risk assessments that pressure-test the system at the key points outlined in this policy advisory opinion.

179. Third, Meta disclosed that one of the categories it uses for Early Response Secondary Review is "historically over-enforced entities." This means that the company has already identified entities where Meta acknowledges it is unable to enforce its policies consistently and effectively. In addition to providing such entities access to over-enforcement mistake-prevention programs, Meta should use this data to inform how to improve its enforcement practices at scale. Meta should measure over-enforcement of these entities and it should use that data to help identify other over-enforced entities. Reducing that metric should be an explicit and high-priority goal for the company.

180. There are additional metrics Meta should develop and monitor to better align mistake-prevention strategies with human rights standards. For example, Meta should establish new metrics to quantify the impact of leaving violating content on the platform. In particular, the company should calculate the number of views a piece of content that is ultimately removed accumulates while pending review because of mistake-prevention mechanisms. Meta should determine a baseline for this metric and report on goals to reduce it.

181. Meta also disclosed that it also takes steps to address some issues related to under-enforcement. These include "classifiers to detect content that likely violates our policies; user reports that identify potentially violating content; human review sweeps where our teams review potentially violating content; High Risk Early Review Operations (HERO), a system where humans review content predicted to go viral; and reporter appeals, where users who report violating content may appeal [Meta's] decision."

182. The Board notes that efforts outside of at-scale automated enforcement and appeals have a narrow scope. Additionally, some of these initiatives compete for resources with cross-check. For example, HERO review is done by the market teams, which also must devote capacity to cross-check. HERO also only affords reviews to content expected to go viral. The Board agrees that high-reach content may cause more harm but believes it should be accompanied by efforts to improve moderation comprehensively. Meta should continue to invest in early detection and warning systems; and hiring and embedding people with local and language expertise in their trust and safety, content review operation, and mistake-prevention system list-creation efforts.

VII. Transparency recommendations

183. The Board has made a series of recommendations about how Meta should design and govern any false-positive mistake-prevention program. Meta's human rights responsibilities also mean it should provide transparency to the public about these programs. Transparency reports should contain comprehensive data so that users and the public understand how the program is functioning and what its consequences on public discourse may be. In addition to the described metrics, the Board recommends Meta include:

a. Overturn rates for false-positive mistake-prevention systems, disaggregated according to design choices and enforcement teams (e.g., Markets, Early Response, contractors, etc.) For example, the Board has recommended that Meta create separate streams for different categories of entities or content based on their expression and risk profile. The overturn rate should be reported for any entity-based and content-based systems, and categories of entities or content included.
b. The total number and percentage of escalation-only policies applied due to false-positive mistake prevention programs relative to total enforcement decisions.

c. Average and median time to final decision for content subject to false-positive mistake-prevention programs, disaggregated by country and language.
d. Aggregate data regarding any lists used for mistake-prevention programs, including the type of entity and region.
e. Rate of erroneous removals (false positives) on all reviewed content, including the total amount of harm generated by these false positives measured as the predicted total views on the content (i.e., over-enforcement).
f. Rate of erroneous keep-up decisions (false negatives) on content, including the total amount of harm generated by these false negatives, measured as the sum of views the content accrued (i.e., under-enforcement).

184. The Board has previously recommended that Meta disclose error rates in general, but also that it should "report on the relative error rates of determinations made through cross check compared with ordinary enforcement procedures." The Board believes that Meta's focus on prevalence, while useful in certain specific contexts, does not provide the right incentives to the company or the right tools for the public to understand how Meta's content moderation ecosystem is functioning.

185. Meta told the Board that it is "currently investing in an aggregate topline metric measurement that allows us to understand false positives through the entire system and are working to build this metric which we hope to share externally in our transparency reporting. This metric would be the counter metric to our false negative measurement that is currently reported through prevalence metrics." This is a step in the right direction and the Board urges Meta to complete this work as soon as possible.

186. In addition to the metrics emphasized in previous sections, which serve both to benchmark improvement and to provide information, Meta should further provide basic information in its Transparency Center regarding the functioning of any mistake-prevention system it uses that identifies entities or users for additional protections. The Board understands the potential for user adversarialism to attempt to bypass enforcement, and Meta may choose to summarize some points of its enforcement practices. The current level of transparency is inadequate and not justified by fear of adversarial risk.

187. More generally, the Board notes that providing greater transparency to external researchers, in particular access to data, is an essential component of oversight for mistake-prevention systems. Throughout the stakeholder engagement conducted for this analysis, the Board heard concern about Meta seeking to limit its current data access programs for external parties. Considering that systems like cross-check require making complex trade-offs, independent researchers could provide Meta with valuable insights on the impacts of its choices. The Board believes Meta should institute a pathway for external researchers to gain access to non-public data about the cross-check system that would allow them to understand the program more fully through public-interest investigations and provide their own recommendations for improvement. While mitigation measures to protect user

privacy must be taken, Meta could and should allow for greater understanding of how its platforms work.

VIII. Annex with recommendations and measures of implementation

The Board made multiple recommendations to Meta in its policy advisory opinion. This annex pairs those recommendations with measures of implementation to monitor Meta's progress. Meta should provide information about its implementation work in its quarterly reports on the Board. Additionally, Meta should convene a biannual meeting of high-level responsible officials to brief the Board on its work to implement the policy advisory opinion recommendations.

#	Recommendation	Measures of implementation
Entity-based mistake-prevention governance		
1	Meta should split, either by distinct pathways or prioritization, any list-based over-enforcement prevention program into separate systems: one to protect expression in line with Meta's human rights responsibilities, and one to protect expression that Meta views as a business priority that falls outside that category.	Meta provides the Board with information detailing how both inclusion and operation are split for these categories of entities. Meta publicizes the details about these systems in its Transparency Center. *Enforcement*
2	Meta should ensure that the review pathway and decision-making structure for content with human rights or public interest implications including its escalation paths, is devoid of business considerations. Meta should take steps to ensure that the team in charge of this system does not report to public policy or government relations teams or those in charge of relationship management with any affected users.	Meta provides the Board with information detailing the decision-making pathways and teams involved in the content moderation of content with human rights or public interest implications. *Enforcement*
3	Meta should improve how its workflow dedicated to meet Meta's human rights responsibilities incorporates context and language expertise on enhanced review, specifically at decision making levels.	Meta provides the Board with information detailing how it has improved upon its current process to include language and context expertise at the moment that context-based decisions and policy exceptions are being considered. *Enforcement*

4	Meta should establish clear and public criteria for list-based mistake-prevention eligibility. These criteria should differentiate between users who merit additional protection from a human rights perspective and those included for business reasons.	Meta releases a report or Transparency Center update detailing the criteria for list-based enhanced review eligibility for the different categories of users the program will enroll. *Transparency*
5	Meta should establish a process for users to apply for over-enforcement mistake-prevention protections should they meet the company's publicly articulated criteria. State actors should be eligible to be added or apply based on these criteria and terms but given no other preference.	Meta implements a publicly and easily accessible, transparent application system for any list-based over-enforcement protection, detailing what purposes the system serves and how the company assesses applications. Meta includes the number of entities that successfully enrolled in mistake-prevention through application, their country and category each year in its Transparency Center. *Enforcement*
6	Meta should ensure that the process for list-based inclusion, regardless of who initiated the process (the entity itself or Meta) involves, at minimum: (1) an additional, explicit, commitment by the user to follow Meta's content policies; (2) an acknowledgement of the program's particular rules; and (3) a system by which changes to the platform's content policies are proactively shared with them.	Meta provides the Board with the complete user experience for onboarding into any list-based system, including how users commit to content policy compliance and how they are notified of policy changes. *Enforcement*
7	Meta should strengthen its engagement with civil society for the purposes of list creation and nomination. Users and trusted civil society organizations should be able to nominate others that meet the criteria. This is particularly urgent in countries where the company's limited presence does not allow it to identify candidates for inclusion independently.	Meta provides information to the Board on how the company engages with civil society to determine list-based eligibility. Meta provides data in its Transparency Center, disaggregated by country, on how many entities are added as a result of civil society engagement as opposed to proactive selection by Meta. *Enforcement*
8	Meta should use specialized teams, independent from	Meta provides the Board with internal documents detailing which teams handle

	political or economic influence, including from Meta's public policy teams, to evaluate entities for list inclusion. To ensure criteria are met, specialized staff, with the benefit of local input, should ensure objective application of inclusion criteria.	list creation and where they sit in the organization. *Enforcement*
9	Meta should require that more than one employee be involved in the final process of adding new entities to any lists for false positive mistake-prevention systems. These people should work on different but related teams.	Meta provides the Board with information detailing the process by which new entities are added to lists, including how many employees must approve inclusion and what teams they belong to. *Enforcement*
10	Meta should establish clear criteria for removal. One criterion should be the amount of violating content posted by the entity. Disqualifications should be based on a transparent strike system, in which users are warned that continued violation may lead to removal from the system and or Meta's platforms. Users should have the opportunity to appeal such strikes through a fair and easily accessible process.	Meta provides the Board with information detailing the threshold of enforcement actions against entities at which their protection under a list-based program is revoked, including notifications sent to users when they receive strikes against their eligibility, when they are disqualified, and their options for appeal. It should also provide the Board with data about how many entities are removed each year for posting violating content. *Enforcement*
11	Meta should establish clear criteria and processes for audit. Should entities no longer meet the eligibility criteria, they should be promptly removed from the system. Meta should review all included entities in any mistake-prevention system at least yearly. There should also be clear protocols to shorten that period where warranted.	Meta provides the Board with data on the amount, type of entity, and reason for removal from entity lists as a result of audits, along with a timeline for conducting audits periodically. *Enforcement*
List transparency		
12	Meta should publicly mark the pages and accounts of entities receiving list-based protection in the following categories: all state actors and political candidates, all business partners, all media actors, and all other public figures included because of the	Meta marks all entities in these categories as benefiting from an entity-based mistake prevention program and announces the change in its Transparency Center. *Transparency*

	commercial benefit to the company in avoiding false positives. Other categories of users may opt to be identified.	
13	Meta should notify users who report content posted by an entity publicly identified as benefiting from additional review that special procedures will apply, explaining the steps and potentially longer time to resolution.	Meta provides the Board with the notifications for users that report content from users identified as benefiting from additional review and confirm global implementation and data that shows these notifications are consistently shown to users. *Enforcement*
14	Meta should notify all entities that it includes on lists to receive enhanced review and provide them with an opportunity to decline inclusion.	Meta provides the Board with (1) the notifications sent to users informing them of their inclusion in a list-based enhanced review program and offering them the option to decline; and Meta (2) publicly reports annual numbers in its Transparency Center on the amount of entities, per country, that declined inclusion. *Enforcement*
Enhanced review and prioritization		
15	Meta should consider reserving a minimum amount of review capacity by teams that can apply all content policies (e.g., the Early Response Team) to review content flagged through content-based mistake-prevention systems.	Meta provides the Board with documentation showing its process of consideration of this recommendation and the rationale for its decision on whether to implement it and publishes this justification to their Transparency Center. *Enforcement*
16	Meta should take measures to ensure that additional review decisions for mistake-prevention systems that delay enforcement are taken as quickly as possible. Investments and structural changes should be made to expand the review teams so that reviewers are available and working in relevant time zones whenever content is flagged for any enhanced human review.	Meta provides the Board with data that demonstrates a quarter-over-quarter reduction in time-to-decision for all content receiving enhanced review, disaggregated by category for inclusion and country. *Enforcement*
17	Meta should not delay all action on content identified as potentially severely violating and	Meta updates its Transparency Center with its new approach to enforcement action during the time when content

	should explore applying interstitials or removals pending any enhanced review. The difference between removal or hiding and downranking should be based on an assessment of harm, and may be based, for example, on the content policy that has possibly been violated. If content is hidden on these grounds, a notice indicating that it is pending review should be provided to users in its place.	receives enhanced review and provides the Board with information detailing the enforcement consequence it will apply based on content-specific criteria. Meta shares with the Board data on the application of these measures and their impact *Enforcement*
Resourcing		
18	Meta should not operate these programs at a backlog. Meta should not, however, achieve gains in relative review capacity by artificially raising the ranker threshold or having its algorithm select less content.	Meta provides the Board with data that demonstrates a quarter-over-quarter reduction in total backlogged content and amount of days with a backlog for cross-check review queues. *Enforcement*
19	Meta should not automatically prioritize entity-based secondary review and make a large portion of the algorithmically selected content-based review dependent on extra review capacity.	Meta provides the Board with internal documents detailing the distribution of review time and volume between entity-based and content-based systems. *Enforcement*
20	Meta should ensure that content that receives any kind of enhanced review because it is important from a human rights perspective, including content of public importance, is reviewed by teams that can apply exceptions and context.	Meta provides the Board with information that shows the percentage of content receiving review by teams that can apply exceptions and context because it has been posted by an entitled entity or because it has been identified algorithmically as meriting enhanced review disaggregated by mistake-prevention system (e.g., GSR vs. ERSR). *Enforcement*
Automatic bars to enforcement ('technical corrections')		

21	Meta should establish clear criteria for the application of any automatic bars to enforcement ('technical corrections'), and not permit such bars for high severity content policy violations. At least two teams with separate reporting structures should participate in granting technical corrections to provide for cross-team vetting.	Meta publishes the number of entities currently benefitting from a "technical correction" on an annual basis, with indication of what content policies are barred from enforcement. *Enforcement*
22	Meta should conduct periodic audits to ensure that entities benefitting from automatic bars to enforcement ('technical corrections') meet all criteria for inclusion. At least two teams with separate reporting structures should participate in these audits to provide for cross-team vetting.	Meta provides information to the Board on its periodic list auditing processes. *Enforcement*
23	Meta should conduct periodic multi-team audits to proactively and periodically search for unexpected or unintentional bars to enforcement that may result from system error.	Meta publishes information annually on any unexpected enforcement bars it has found, their impact, and the steps taken to remedy the root cause. *Enforcement*
Procedural fairness		
24	Meta should ensure that all content that does not reach the highest level of internal review is able to be appealed to Meta.	Meta publishes information on the number of content decisions made through enhanced review pathways that were not eligible for appeal. This yearly data, disaggregated by country, should be broken down in a way that explains what, if any, percentage of the content did not get an appeal because it reached global leadership review. *Enforcement*
25	Meta must guarantee that it is providing an opportunity to appeal to the Board for all content the Board is empowered to review under its governing documents, regardless of whether the content reached the highest levels of review within Meta.	Meta publicly confirms all content covered under the Board's governing documents are receiving Oversight Board appeal IDs to submit a complaint to the Board, providing documentation to demonstrate where steps have been taken to close appeal availability gaps. Meta creates an accessible channel for users to achieve

		prompt redress when they do not receive an Oversight Board appeal ID. *Enforcement*
Learning and improvement		
26	Meta should use the data it compiles to identify "historically over-enforced entities" to inform how to improve its enforcement practices at scale. Meta should measure over-enforcement of these entities and it should use that data to help identify other over-enforced entities. Reducing over-enforcement should be an explicit and high-priority goal for the company.	Meta provides data to the public that shows quarter-over-quarter declines in over-enforcement and documentation that shows that the analysis of content from "historically over-enforced" entities is being used to reduce over-enforcement rates more generally. *Enforcement*
27	Meta should use trends in overturn rates to inform whether to default to the original enforcement within a shorter time frame or what other enforcement action to apply pending review. If overturn rates are consistently low for particular subsets of policy violations or content in particular languages, for example, Meta should continually calibrate how quickly and how intrusive an enforcement measure it should apply.	Meta provides the Board with data detailing the rates at which queued content remains up or is taken down, broken out by country, policy area, and other relevant metrics, and describes changes made on an annual basis. *Enforcement*
Improving program accountability		
28	Meta should conduct periodic reviews of different aspects of its enhanced review system, including content with the longest time to resolution and high-profile violating content left on platform.	Meta publishes the results of reviews to the cross-check system on an annual basis, including summaries of changes made as a result of these reviews. *Transparency*
29	Meta should publicly report on metrics that quantify the adverse effects of delayed enforcement as a result of enhanced review systems, such as views accrued on content that was preserved on the platform as a result of mistake-prevention systems but	Meta includes one or more key metrics demonstrating the negative consequences of delayed enforcement pending enhanced review mechanisms in the Community Standards Enforcement Report, along with goals to reduce these metrics and progress in meeting those goals.

	was subsequently found violating. As part of its public reporting, Meta should determine a baseline for these metrics and report on goals to reduce them.	*Transparency*
30	Meta should publish regular transparency reporting focused specifically on delayed-enforcement false-positive prevention systems. Reports should contain data that permits users and the public to understand how these programs function and what their consequences on public discourse may be. At minimum, the Board recommends Meta include: a. Overturn rates for false-positive mistake-prevention systems, disaggregated according to different factors. For example, the Board has recommended that Meta create separate streams for different categories of entities or content based on their expression and risk profile. The overturn rate should be reported for any entity-based and content-based systems, and categories of entities or content included. b. The total number and percentage of escalation-only policies applied due to false-positive mistake-prevention programs relative to total enforcement decisions. c. Average and median time to final decision for content subject to false-positive mistake-prevention programs, disaggregated by country and language.	Meta releases annual transparency reporting including these metrics. *Transparency*

	d. Aggregate data regarding any lists used for mistake-prevention programs, including the type of entity and region. e. Rate of erroneous removals (false positives) versus all reviewed content, including the total amount of harm generated by these false positives measured as the predicted total views on the content (i.e., over-enforcement) f. Rate of erroneous keep-up decisions (false negatives) on content, including the total amount of harm generated by these false positives, measured as the sum of views the content accrued (i.e., under-enforcement)	
31	Meta should provide basic information in its Transparency Center regarding the functioning of any mistake-prevention system it uses that identifies entities or users for additional protections.	A section is added to the Transparency Center explaining its array of mistake-prevention systems (the Board understands the potential for user adversarialism to attempt to bypass enforcement, and Meta may choose to summarize some points of its enforcement practices). *Transparency*
32	Meta should institute a pathway for external researchers to gain access to non-public data about false-positive mistake-prevention programs that would allow them to understand the program more fully through public-interest investigations and provide their own recommendations for improvement. The Board understands that data privacy concerns should require stringent vetting and data aggregation.	Meta discloses a pathway for external researchers to obtain non-public data on false positive mistake-prevention programs. *Transparency*

PX0334

Meta

Language

Transparency Center

Home → Policies → How Meta improves

Proactive rate

UPDATED FEB 22, 2023

This metric shows the percentage of all content or accounts acted on that we found and actioned before users reported them to us. We use this metric as an indicator of how effectively we detect violations.

Our investments in machine learning technology are critical to help us detect faster.

We balance machine learning with a trained team of experts who review and take action on violating content.

The rate at which we can proactively detect potentially violating content is high for some violations, meaning we find and action most content before users report it to us. This is especially true where we've been able to build machine learning technology that automatically identifies content that might violate our standards.

Such technology is very promising but is still years away from being effective for all kinds of violations. For example, there are still limitations in the ability to understand context and nuance, especially for text-based content. This creates additional challenges for proactively detecting certain violations.

The metric can go up or down due to external factors. For example, a cyberattack during which spammers share 10 million posts featuring the same malicious URL. If we detected the malicious URL before any user reported it to us, proactive rate would go up during the cyberattack and go down afterward, even if our detection technology didn't change during the period. This metric can also increase or decrease based on how our processes and tools change — for example, it might go up if our detection technology gets better, but it might go down if our user reporting improves and we rely less on proactive detection.

Since this metric is based on the amount of content actioned, many of the same considerations apply. Our proactive rate doesn't reflect how long it takes us to detect violating content or how many times it was viewed before detection. It also doesn't reflect how many violations we failed to detect altogether or how many times that content was viewed. And though the percentage of content we proactively detect can be very high — as high as 99% in some categories — even the remaining small percentage can cause significant impact to people.

How we calculate our proactive rate

We calculate this percentage as: the number of pieces of content acted on that we found and actioned before people using Facebook or Instagram reported them, divided by the total number of pieces of content we took action on.

For fake accounts on Facebook, we calculate this metric as the percentage of Facebook accounts disabled for being fake that we found and actioned before users reported them to us. It's calculated as the number of disabled accounts we found and actioned before users reported them, divided by the total number of accounts disabled for being fake.

Caveats

We compute our proactive rate using a strict attribution of user reports to content. For example, if someone reports a Page and, while reviewing the Page, we identify and act on some violating content within that Page, we would report that we had actioned that content proactively (unless there were specific additional user reports of it).

View the latest Community Standards Enforcement Report

NEXT

Appealed content

PREVIOUS

Content actioned

Meta

POLICIES

ENFORCEMENT

SECURITY

FEATURES

GOVERNANCE

REPORTS

RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0336

Meta

Transparency Center

Instagram

Download (CSV)

Overview

Adult Nudity and Sexual Activity

Bullying and Harassment

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Dangerous Organizations: Terrorism and Organized Hate

Fake Accounts

Hate Speech

Restricted Goods and Services

Spam

Suicide and Self-Injury

Violence and Incitement

Violent and Graphic Content

Spam

We do not allow spam on Facebook. Spam is a broad term used to describe content that is designed to be shared in deceptive and annoying ways or attempts to mislead users to drive engagement. Spam spreads in a number of ways: It can be automated (published by bots or scripts) or coordinated (when an actor uses multiple accounts to spread deceptive content).

Spammers aim to build audiences to inflate their content's distribution and reach, typically for financial gain. The tactics spammers use, and our ability to detect them, drive the amount of content we take action on as well as our proactive rate.

Read the policy details

PREVALENCE

How prevalent were spam violations?

We cannot estimate this metric right now. We are working on new methods to measure the prevalence of spam on Instagram. Our existing methods for measuring prevalence, which rely on people to manually review samples of content, do not fully capture this type of highly adversarial violation, which includes deceptive behavior as well as content. Spammy behavior cannot always be detected by reviewing the content alone. We are working on ways to review and classify spammers' behavior to build a comprehensive picture.

CONTENT ACTIONED

How much spam content did we take action on?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as

we confirm accuracy and meaningful data.

How we calculate it Read about this data

PROACTIVE RATE

Of the violating content we actioned for spam, how much did we find and action before people reported it?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as we confirm accuracy and meaningful data.

How we calculate it Read about this data

Correcting mistakes

People can appeal our decisions, unless there are extreme safety concerns. We restore content we incorrectly removed or when circumstances change. Restores can happen from appeals or when we identify issues ourselves.

APPEALED CONTENT

How much of the content we actioned for spam did people appeal?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as we confirm accuracy and meaningful data.

How we calculate it ⓘ **Read about this data**

RESTORED CONTENT

How much actioned content for spam was later restored?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as we confirm accuracy and meaningful data.

How we calculate it **Read about this data**

NOTE:

Due to a temporary reduction in our review capacity as a result of COVID-19, we could not always offer people the option to appeal. We still gave people the option to tell us they disagreed with our decision, which helped us review many of these instances and restore content when appropriate. Starting in Q2 2022, we updated our methodology for how we count appeals to include all instances where content was submitted for additional review, including after people told us that they disagreed with our decision.

NEXT

Suicide and Self-Injury

PREVIOUS

Restricted Goods and Services

Meta

POLICIES
ENFORCEMENT
SECURITY
FEATURES
GOVERNANCE
REPORTS
RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0337

Meta

Transparency Center

Instagram

Download (CSV)

Overview

Adult Nudity and Sexual Activity

Bullying and Harassment

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Dangerous Organizations: Terrorism and Organized Hate

Fake Accounts

Hate Speech

Restricted Goods and Services

Spam

Suicide and Self-Injury

Violence and Incitement

Violent and Graphic Content

Fake Accounts

Our goa is to remove as many fake accounts on Facebook as we can. These include accounts created with malicious intent to violate our policies and personal profiles created to represent a business, organization or non-human entity, such as a pet.

We prioritize enforcement against fake accounts that seek to cause harm. Many of these accounts are used in spam campaigns and are financially motivated.

We expect the number of accounts we action to vary over time due to the unpredictable nature of adversarial account creation. Our detection technology helps us block millions of attempts to create fake accounts every day and detect millions more, often within minutes after creation. We do not include blocked attempts in the metrics we report here.

Read the policy details

PREVALENCE

How prevalent were fake account violations?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as we confirm accuracy and meaningful data.

How we calculate it **Read about this data**

ACCOUNTS ACTIONED

How many fake accounts did we take action on?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as we confirm accuracy and meaningful data.

How we calculate it ⓘ **Read about this data**

PROACTIVE RATE

Of the violating accounts we actioned, how many did we find and action before people reported them?

We cannot estimate this metric right now

We will continue to expand the measurement to more areas as we confirm accuracy and meaningful data.

How we calculate it ⓘ Read about this data

NEXT

Hate Speech

PREVIOUS

Dangerous Organizations: Terrorism and Organized Hate

Meta

POLICIES
ENFORCEMENT
SECURITY
FEATURES
GOVERNANCE
REPORTS
RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0338

Meta

Transparency Center

Instagram

Download (CSV)

Overview

Adult Nudity and Sexual Activity

Bullying and Harassment

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Dangerous Organizations: Terrorism and Organized Hate

Fake Accounts

Hate Speech

Restricted Goods and Services

Spam

Suicide and Self-Injury

Violence and Incitement

Violent and Graphic Content

Adult Nudity and Sexual Activity

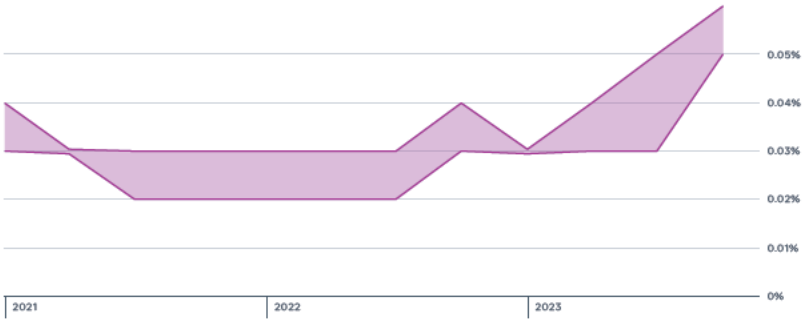
We restrict the display of adult nudity and sexual activity on Facebook and Instagram. We make some exceptions when it is clear the content is being shared in the context of a protest, for educational or medical reasons or a similar reason. On the other hand, we default to removing sexual imagery to prevent non-consensual or underage content from being shared.

This report does not include metrics related to our separate policy on the promotion of sexual assault, violence or exploitation.

Read the policy details

PREVALENCE

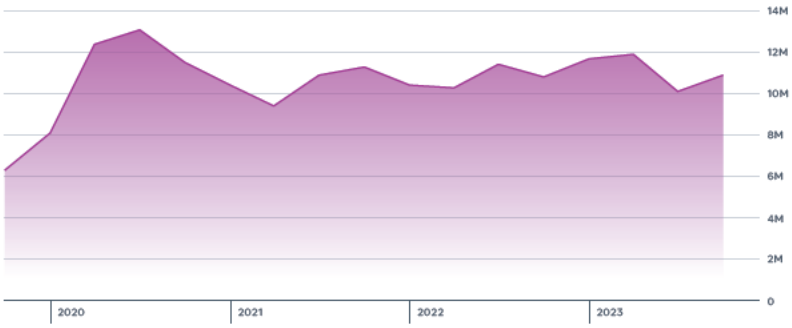
How prevalent were adult nudity and sexual activity violations?



How we calculate it ⓘRead about this data

CONTENT ACTIONED

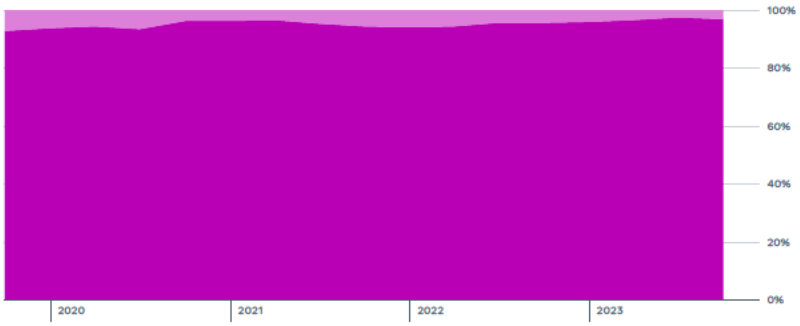
How much adult nudity and sexual activity content did we take action on?



How we calculate it **Read about this data**

PROACTIVE RATE

Of the violating content we actioned for adult nudity and sexual activity, how much did we find and action before people reported it?



Found and actioned by us

Reported by users

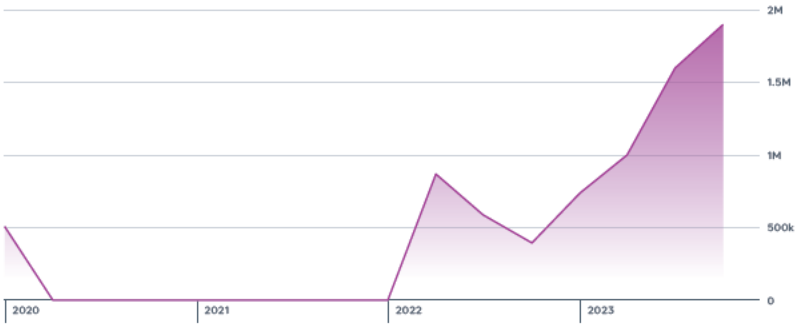
How we calculate it **Read about this data**

Correcting mistakes

People can appeal our decisions, unless there are extreme safety concerns. We restore content we incorrectly removed or when circumstances change. Restores can happen from appeals or when we identify issues ourselves.

APPEALED CONTENT

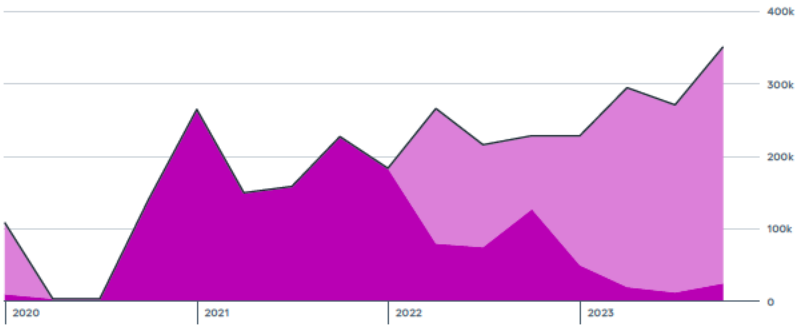
How much of the content we actioned for adult nudity and sexual activity did people appeal?



How we calculate it **Read about this data**

RESTORED CONTENT

How much actioned content for adult nudity and sexual activity was later restored?



Restored without appeal

Restored after appeal

Total

How we calculate it Read about this data

NOTE:

Due to a temporary reduction in our review capacity as a result of COVID-19, we could not always offer people the option to appeal. We still gave people the option to tell us they disagreed with our decision, which helped us review many of these instances and restore content when appropriate. Starting in Q2 2022, we updated our methodology for how we count appeals to include all instances where content was submitted for additional review, including after people told us that they disagreed with our decision.

NEXT

Bullying and Harassment

PREVIOUS

Overview

Meta

POLICIES
ENFORCEMENT
SECURITY
FEATURES
GOVERNANCE
REPORTS
RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0339

Meta

Transparency Center

Instagram

Download (CSV)

Overview

Adult Nudity and Sexual Activity

Bullying and Harassment

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Dangerous Organizations: Terrorism and Organized Hate

Fake Accounts

Hate Speech

Restricted Goods and Services

Spam

Suicide and Self-Injury

Violence and Incitement

Violent and Graphic Content

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

We do not allow content that endangers children, such as content that contains nudity or physical abuse or content that sexually exploits children on Facebook and Instagram. When we find this type of violating content, we remove it, regardless of the context or the person's motivation for sharing it. We may also disable the account of the person who shared it, unless it appears the intent was not malicious (for example, to spread awareness of child exploitation).

We report apparent child exploitation to the National Center for Missing and Exploited Children (NCMEC), a nonprofit that refers cases to law enforcement globally, in compliance with US law. We choose to remove content depicting non-sexualized child nudity to reduce the potential for abuse of the content by others.

Read the policy details

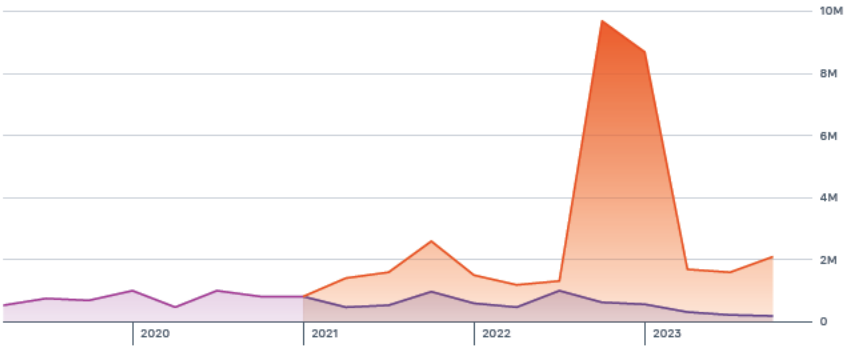
PREVALENCE

How prevalent were child endangerment violations?

We are working on estimating prevalence for child endangerment violations. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.

CONTENT ACTIONED

How much child endangerment content did we take action on?



Child Nudity and Sexual Exploitation

Child Nudity and Physical Abuse

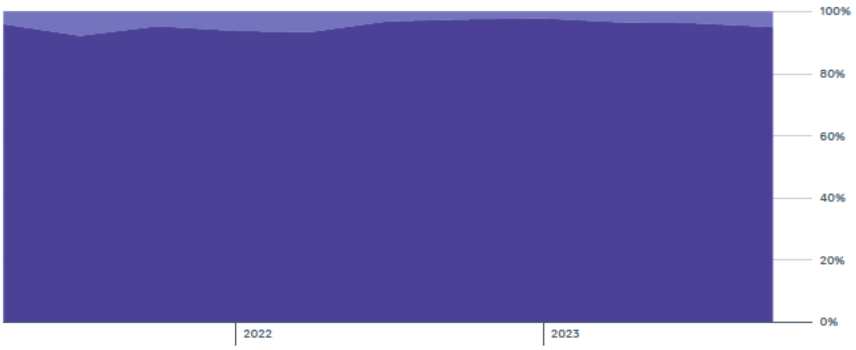
Child Sexual Exploitation

How we calculate it **Read about this data**

PROACTIVE RATE

Of the violating content we actioned for child endangerment, how much did we find and action before people reported it?

Child Nudity and Physical Abuse



Found and actioned by us | Reported by users

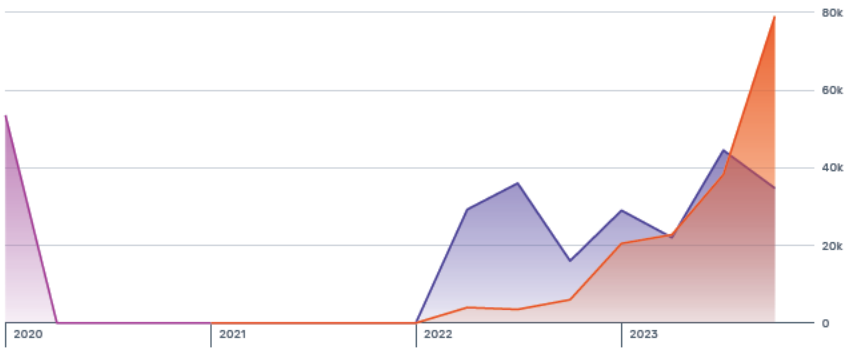
How we calculate it **Read about this data**

Correcting mistakes

People can appeal our decisions, unless there are extreme safety concerns. We restore content we incorrectly removed or when circumstances change. Restores can happen from appeals or when we identify issues ourselves.

APPEALED CONTENT

How much of the content we actioned for child endangerment did people appeal?



Child Nudity and Sexual Exploitation

Child Nudity and Physical Abuse

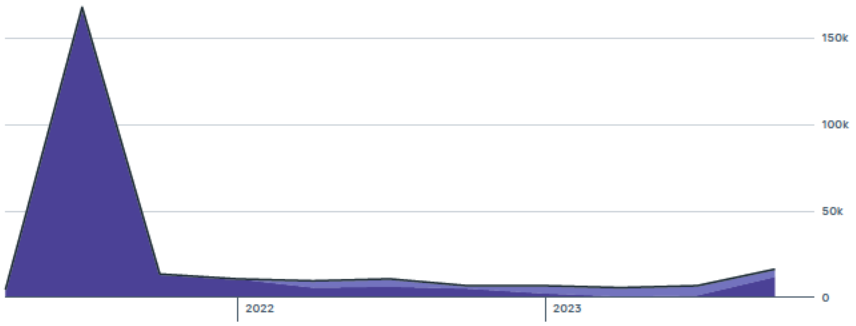
Child Sexual Exploitation

How we calculate it Read about this data

RESTORED CONTENT

How much actioned content for child endangerment was later restored?

Child Nudity and Physical Abuse



Restored without appeal

Restored after appeal

Total

How we calculate it Read about this data

NOTE:

Due to a temporary reduction in our review capacity as a result of COVID-19, we could not always offer people the option to appeal. We still gave people the option to tell us they disagreed with our decision, which helped us review many of these instances and restore content when appropriate. Starting in Q2 2022, we updated our methodology for how we count appeals to include all instances where content was submitted for additional review, including after people told us that they disagreed with our decision.

NEXT

Dangerous Organizations: Terrorism and Organized Hate

PREVIOUS

Bullying and Harassment

Meta

POLICIES

ENFORCEMENT

SECURITY

FEATURES

GOVERNANCE

REPORTS

RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0340

Meta

Transparency Center

Instagram

Download (CSV)

Overview

Adult Nudity and Sexual Activity

Bullying and Harassment

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Dangerous Organizations: Terrorism and Organized Hate

Fake Accounts

Hate Speech

Restricted Goods and Services

Spam

Suicide and Self-Injury

Violence and Incitement

Violent and Graphic Content

Bullying and Harassment

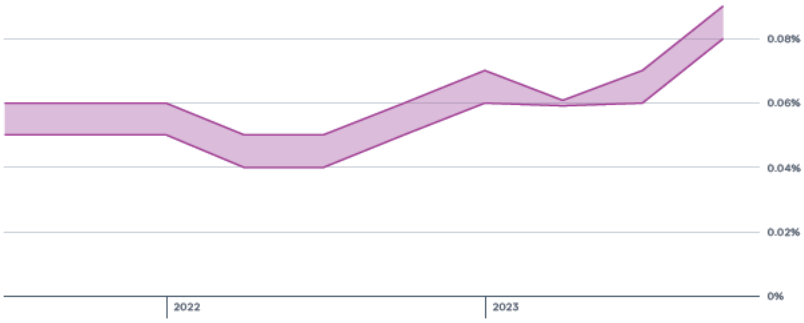
We do not tolerate bullying and harassment on Facebook and Instagram. Because we recognize bullying can be especially harmful for minors, our policies provide heightened protections for them. We want to allow for open and vital discussion of people who are in the news or who have a large public audience, so we do permit more open or critical discourse towards public figures than private individuals.

Because bullying and harassment is highly personal by nature, using technology to proactively detect these behaviors can be more challenging than other types of violations. That's why we also rely on people to report this behavior to us so we can identify and remove it. When measuring prevalence in this area, the metric captures only bullying and harassment where a deeper understanding of context or meaning is not necessary to determine if it violates our policy. We continue to invest in our proactive detection technology to ensure we are tackling the problem and protecting our community.

Read the policy details

PREVALENCE

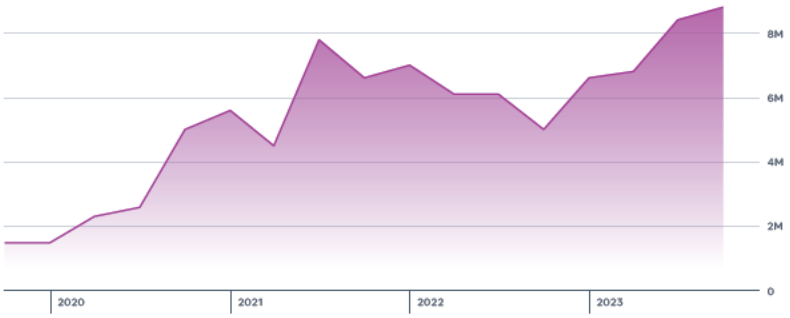
How prevalent were bullying and harassment violations?



How we calculate it ⓘ**Read about this data**

CONTENT ACTIONED

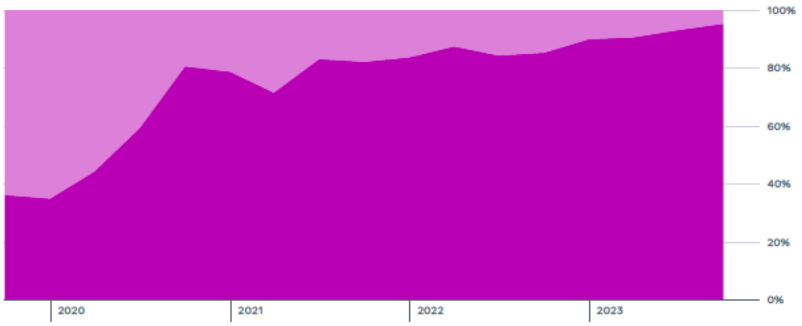
How much bullying and harassment content did we take action on?



How we calculate it **Read about this data**

PROACTIVE RATE

Of the violating content we actioned for bullying and harassment, how much did we find and action before people reported it?



Found and actioned by us

Reported by users

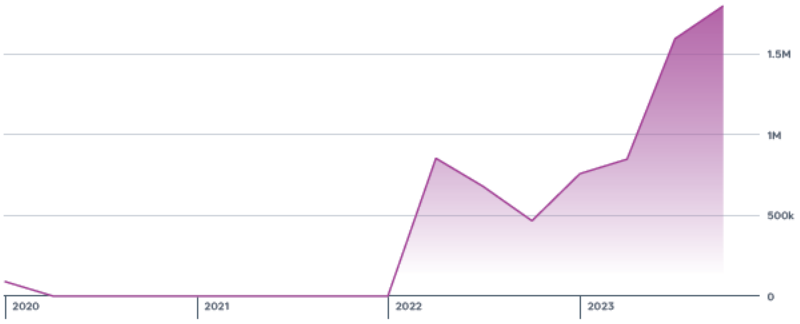
How we calculate it **Read about this data**

Correcting mistakes

People can appeal our decisions, unless there are extreme safety concerns. We restore content we incorrectly removed or when circumstances change. Restores can happen from appeals or when we identify issues ourselves.

APPEALED CONTENT

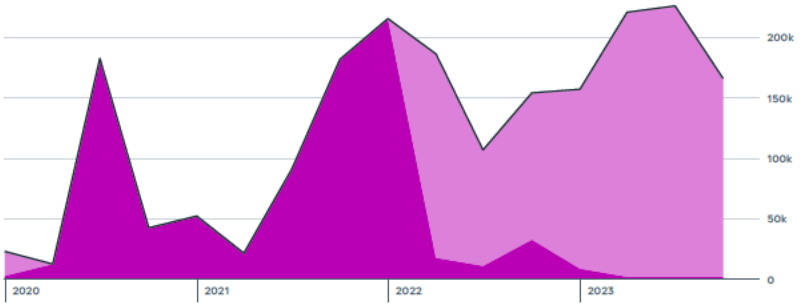
How much of the content we actioned for bullying and harassment did people appeal?



How we calculate it **Read about this data**

RESTORED CONTENT

How much actioned content for bullying and harassment was later restored?



Restored without appeal

Restored after appeal

Total

How we calculate it Read about this data

NOTE:

Due to a temporary reduction in our review capacity as a result of COVID-19, we could not always offer people the option to appeal. We still gave people the option to tell us they disagreed with our decision, which helped us review many of these instances and restore content when appropriate. Starting in Q2 2022, we updated our methodology for how we count appeals to include all instances where content was submitted for additional review, including after people told us that they disagreed with our decision.

NEXT

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

PREVIOUS

Adult Nudity and Sexual Activity

Meta

POLICIES
ENFORCEMENT
SECURITY
FEATURES
GOVERNANCE
REPORTS
RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0347

Meta

Transparency Center

Facebook

Download (CSV)

Overview

Adult Nudity and Sexual Activity

Bullying and Harassment

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Dangerous Organizations: Terrorism and Organized Hate

Fake Accounts

Hate Speech

Restricted Goods and Services

Spam

Suicide and Self-Injury

Violence and Incitement

Violent and Graphic Content

Community Standards Enforcement Report

We want Facebook and Instagram to be places where people have a voice. To create conditions where everyone feels comfortable expressing themselves, we must also protect their safety, privacy, dignity and authenticity. This is why we have the Facebook Community Standards and Instagram Community Guidelines, which define what is and is not allowed in our community.

Q4 2023 report

We publish the Community Standards Enforcement Report on a quarterly basis to more effectively track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive.

What's new

In this February 2024 quarterly report, we share updated metrics for the reporting period from October to December 2023, detailing our progress on content that violates our policies.

Facebook and Instagram policies

Facebook and Instagram share content policies. Content that is considered violating on Facebook is also considered violating on Instagram. Throughout this report, we link to our Community Standards, which include the most comprehensive descriptions of these policies.

As we improve our methodologies, measure violations in more languages or across new parts of Facebook and Instagram and update our policies, the way we define and measure enforcement may change. As a result, historical comparisons may be imperfect.

Learn more

- Review EY's independent, third-party assessment of our Community Standards Enforcement Report from the fourth quarter of 2021.
- Review a report from independent academic experts on their findings and recommendations on our data transparency efforts.

Key Takeaways for Q4 2023

In Q4, we continue to make progress on removing content that violates our Community Standards

- Prevalence of violating content remained relatively consistent across a wide range of violation areas.
- An increase in actioned content for Adult Nudity and Sexual Activity, due to updates to our proactive detection technology that improved accuracy on Live Videos.
- An increase in actioned content and appeals for Dangerous Organizations and Individuals and Violent and Graphic Contents due to the Israel-Hamas war.
- An increase in actioned content, on both Facebook and Instagram, for violations of our Regulated Goods and Services policy, largely related to fireworks and firearms.
- A decrease in appealed content, related to Bullying and Harassment, due to accuracy improvements in our proactive detection technology.

14
Policies on Facebook

12
Policies on Instagram

Read our post about this report

Data by policy area

VIOLENCE AND CRIMINAL BEHAVIOR

Dangerous Organizations: Terrorism and Organized Hate

Restricted Goods and Services: Drugs and Firearms

Violence and Incitement

SAFETY

Suicide and Self-Injury

Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation

Bullying and Harassment

OBJECTIONABLE CONTENT

Hate Speech

Violent and Graphic Content

Adult Nudity and Sexual Activity

INTEGRITY AND AUTHENTICITY

Fake Accounts

Spam

Learn about our measurements

DEEP DIVE

Prevalence
Content actioned
Proactive rate
Appealed content
Restored content
Getting better at measurement
Corrections and adjustments

NEXT

Adult Nudity and Sexual Activity

Meta

POLICIES

ENFORCEMENT

SECURITY

FEATURES

GOVERNANCE

REPORTS

RESEARCH TOOLS

Privacy Policy Terms of Service Cookies

PX0348

WhatsApp

Stopping Abuse:

How WhatsApp Fights Bulk Messaging and Automated Behavior

February 6, 2019

Contents

Summary 3
Introduction 4
WhatsApp Design 5
Detecting and Tackling Abuse 5
At Registration 6
While Messaging 7
In Response to Negative Feedback. 9
Leading Advances with Machine Learning. 10
Addressing Challenges Beyond Our Product 10
Privacy and Safety 11

Summary

From the early days of email to today's social networks, internet platforms have long had to confront abuse including bulk messaging and automated behavior that spreads problematic content. WhatsApp takes a multi-pronged approach to this challenge while also working to maintain the private nature of our service. We have built sophisticated machine learning systems to detect abusive behavior and ban suspicious accounts at registration, during messaging, and in response to user reports. We remove over two million accounts per month for bulk or automated behavior — over 75% without a recent user report. These efforts are particularly important during elections where certain groups may attempt to send messages at scale. While there are many actors trying to abuse the free service we provide, we are constantly advancing our anti-abuse operations to keep our platform safe.

Introduction

WhatsApp was built for private messages: to help people chat with their loved ones, conduct business, or talk confidentially with a doctor. Instead of encouraging users to build an audience and share widely, WhatsApp is designed to help people share with others they know or get helpful information from a business. To protect the privacy and security of our users, WhatsApp provides end-to-end encryption by default, which means only the sender and recipient can see the content of messages.

Our service is not a broadcast platform. We place limits on group sizes and how users send messages. Approximately 90% of the messages sent on WhatsApp are from one person to another, and the majority of groups have fewer than ten people. WhatsApp requires the message sender to know the phone number of the recipient. Whenever a user receives a message from someone not in their address book, we ask the user if they want to block or report the sender.

As with any communications platform, sometimes people attempt to exploit our service. Some may want to distribute click-bait links designed to capture personal information, while others want to promote an idea. Regardless of the intent, automated and bulk messaging violates our terms of service and one of our priorities is to prevent and stop this kind of abuse.

This paper explains our approach. We have had to balance our desire for transparency while not undermining the effectiveness of these anti-abuse measures. Given the nature of end-to-end encryption, we focus on the behavior of accounts and user-reported content[1]. We employ a team of engineers, data scientists, analysts, and researchers with a wide range of backgrounds and expertise to deter and prevent abuse. We pay close attention to user feedback and engage with experts in misinformation, cybersecurity, and election integrity. Through this comprehensive effort we seek to make it difficult for adversaries to abuse the platform we created to connect people.

[1]WhatsApp scans unencrypted information like profile photos for child exploitative imagery. Please see this FAQ for more.

WhatsApp Design

While many online services require an email to sign up, WhatsApp is based on phone numbers and requires that all users have one. This presents unique opportunities for us to catch those seeking to abuse our service.

Mobile numbers are activated through SIM cards, which are distributed by carriers. Buying a SIM card is often simple for a normal user, but SIM cards can be logistically difficult to acquire at scale. In some places, identification is required for purchase; costs may vary across geographies. We design our abuse detection with these factors in mind. Phone numbers originating from areas with a history of fraud may indicate to WhatsApp there is a problem with an account. It can also be a signal to users because receiving a message from an unfamiliar number often indicates something is not right.

Of course, other types of messaging services see abuse as well. A recent analysis of U.S. communications services cited by the CTIA indicated that 2.8% of SMS messages and 50% of e-mail are spam. As on these platforms, preventing users from receiving these kinds of messages is a top priority. Since turning on end-to-end encryption by default in 2016, WhatsApp has worked to reduce spam to maintain the health of our service.

Sending mass messages from a mobile phone is time consuming. It's why people attempting to send bulk messages often build hardware or software to try and automate the process. However, those steps pose financial cost and risk. Our technology aims to identify accounts using automation and take action to stop them. In so doing, we aim to raise the costs associated with this abuse and deter bad actors from trying in the future.

Detecting and Tackling Abuse

Attempting to distribute content en masse on WhatsApp requires users to work around the design of our platform. Doing so exhibits signals that we use to identify abusive accounts. When we are confident that an account is abusive, we ban the account from WhatsApp altogether. We then use the information available to us to reverse engineer past behavior to prevent similar abuse in the future.

Our goal is to identify and stop abusive accounts as quickly as possible, which is why identifying these accounts manually is not realistic. Instead, we have advanced machine learning systems that take action to ban accounts, 24 hours a day, 7 days a week.

This abuse detection operates at three stages of an account's lifestyle: at registration; during messaging; and in response to negative feedback, which we receive in the form of user reports and blocks. A team of analysts augments these systems to evaluate edge cases and help improve our effectiveness over time. We continually work to detect abuse as early as possible to have the greatest impact at preventing abuse before it can reach people.

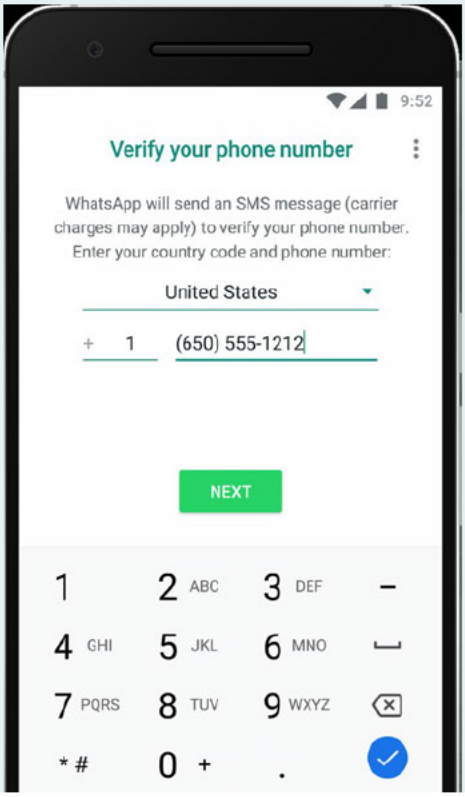
Over the last three months, WhatsApp banned over two million accounts per month for bulk or automated behavior and over 75% of those accounts did not have any recent user reports. To address rare cases when we ban an account in error, we also have a team that reviews and responds to user appeals. In that process, users may be asked for additional information in order to regain access to their account.

2 million
Accounts banned monthly

1.5 billion
Monthly active users

At Registration

All accounts must pass through a common checkpoint: registration. This is the stage where WhatsApp sends a temporary code via SMS or a phone call. Users must enter this one-time code to prove they have control over the phone number and SIM card for their account. We also provide optional two-step verification to prevent attackers from attempting to take over accounts they do not actually own. Since all accounts must go through registration before being able to send any messages, it is an effective starting point to concentrate our efforts.



WhatsApp cares deeply about the privacy of our users, which is why we do not ask for much information from them. Yet even basic account information along with an IP address and associated carrier information can be used to teach our machine learning systems the difference between bulk and normal registrations.

Because we ban accounts that send a high volume of messages, coordinated campaigns often try to spread their activity across many different accounts. We therefore work to understand the behavioral cues indicating bulk registrations. For example, our systems can detect if a similar phone number has been recently abused or if the computer network used for registration has been associated with suspicious behavior. As a result, we're able to detect and ban many accounts before they register — preventing them from sending a single message. In the same three month period, roughly 20% of account bans happened at registration time.

over 75%

accounts banned without a recent user report

roughly 20%

accounts banned at registration

While Messaging

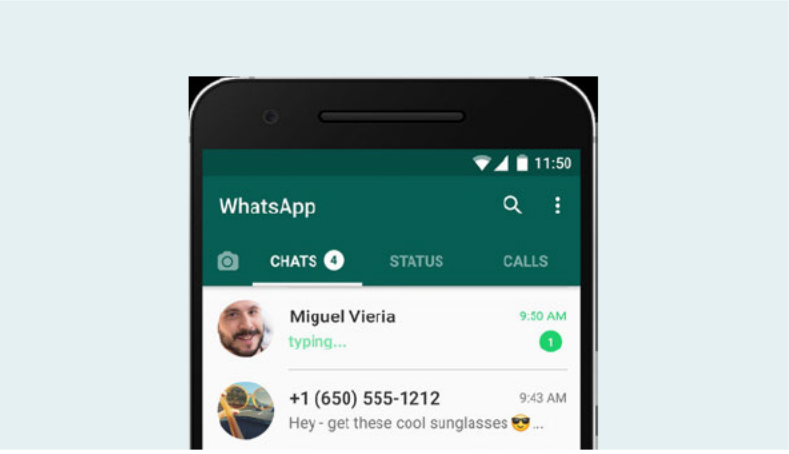
For accounts that complete registration, we evaluate how they behave in real time. All messages are end-to-end encrypted, which means that WhatsApp cannot see the content of messages passing through our system, though we are able to analyze the frequency of account activity.

Normal users operate relatively slowly on WhatsApp, tapping messages one at a time or occasionally forwarding content. The intensity of user activity can provide a signal that accounts are abusing WhatsApp.

For example, an account that registered five minutes before attempting to send 100 messages in 15 seconds is almost certain to be engaged in abuse, as is an account that attempts to quickly create dozens of groups or add thousands of users to a series of existing groups. We ban these accounts immediately and automatically. In less-obvious situations, a new account might message dozens of recipients who do not have the sender's account in their contacts. This could be the beginning of a spam attack, or it could be an innocent user simply telling their contacts about a new phone number. In these cases, we consider historical information (for instance, how suspicious their registration was) in order to separate abnormal — but innocuous — user behavior. In sum, our detection systems evaluate hundreds of factors to shut down abuse.

We also build detection mechanisms that directly target automation. For example, we display at the top of a chat thread when a user is typing. Spammers attempting to automate messaging may lack the technical ability to forge this typing indicator. If an account continually sends messages without triggering the typing indicator, it can be a signal

of abuse, and we will ban the account. This is just one example of the combinations of actions or inactions that our artificial intelligence automatically analyzes to make decisions on which accounts to ban.



We have advanced our technology to the point where we combine signals to make rapid determinations without human intervention. We learn from how other users have behaved and surmise the reputation of new accounts. For example, if an active computer network has recently been used by known abusers, we have more reason to believe a new account on that network is likely to be abusive. These behaviors, and other signals about current and past behavior, are helpful in finding coordinated campaigns that are trying to abuse WhatsApp.

It is important that we try and understand the diverse motivations at hand. For example, someone that has an economic incentive may want to encourage unsuspecting users to tap on a suspicious link, which could lead to a website that harvests personal information from a user. The fact we ban accounts quickly drives up their costs so that any incremental revenue gained through such tactics may not be worth the expense to try in the first place.

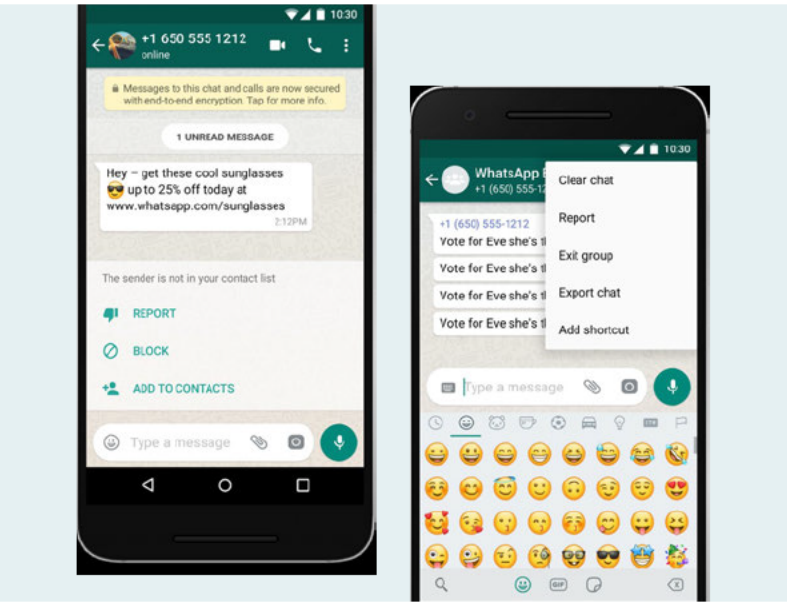
The goals of someone with a political motivation, such as attempting to distort the public discussion with false information, could be different. They seek to organize and drive interest within a particular country, region, or even city - often within a limited time window. To do so they may acquire targeted lists of phone numbers from third parties to message users over WhatsApp without their consent. For this reason, we maintain limits on how many groups an account can create within a certain time period and ban accounts with suspicious group behavior, if appropriate, even if their activity rates are low or have yet to demonstrate high reach. Whenever a user gets added to a group from someone outside their contact list we display an option that asks if the user wants to “report” or “exit” the group. Given the increase concern about abusive activities from politically motivated actors, this is an area of deep focus for our team.

In Response to Negative Feedback

Finally, if an account accumulates negative feedback, such as when other users submit reports or block the account, our systems evaluate the account and take appropriate action.

WhatsApp makes it easy for users to report a problem, and we encourage users to do so. Whenever a user receives a message for the first time from an unknown number, we immediately display options that enable them to "report" or "block" the sender's account. In addition, users can report a problem to WhatsApp at any time in an individual or group chat.

When a user sends a report, our machine learning systems review and categorize the reports so we can better understand the motivation of the account, such as whether they are trying to sell a product or seed misinformation. This new information not only helps us ban accounts, but it also helps us improve our machine learning systems to take earlier action in the future (such as at registration or while messaging, as described above).



While WhatsApp seeks to quickly ban abuse, we recognize the reporting tool can be manipulated. Suppose one group of actors wants to prevent an individual from using WhatsApp. They might organize other users to make a series of reports against their victim. If our systems looked only at report rates and banned any accounts exceeding a threshold, attacks like this could knock legitimate users off our platform. Instead, we evaluate the relationship between the reporter and the reported user to rapidly evaluate the behavior. For instance, when someone reports the first message they receive from an unknown number, we have higher confidence that the report is legitimate. In cases when the reported number did not initiate the communication, we work to ensure there was no coordination among others to falsely report the account.

Leading Advances with Machine Learning

We are constantly working to stay ahead of highly-motivated abusers. As WhatsApp has become more popular we have expanded our detection efforts. The signals we analyze have become complex enough that they're too numerous for an individual to evaluate on a case-by-case basis. Fortunately, computers are very good at considering a large number of factors at once. To that end, WhatsApp has developed machine learning models to help spot abuse quickly.

Machine learning models require basic components known as features, labels, and infrastructure to complete their tasks. Features are the specific signals — such as number of reports, rate of messaging, reputation of other users sharing the same computer network, and so on — that provide insight into abusive behavior. We use labels to mark the worst offenders and distinguish between their behaviors from those of regular users. We use the features and labels to teach our systems to better predict whether a user is likely to be banned in the future. For example, if they determine that an account's behavior at registration matches the behavior of other accounts that were banned, we will ban the account even before it can send any messages. If these systems previously caught accounts after they sent 50 messages, a machine learning model trained on the features and labels from those accounts will catch similar attempts much faster.

Finally, we need infrastructure to train these classifiers, evaluate their performance, and analyze the sending behavior of actual users. Here we worked to adapt the so-called "Facebook Immune System" for use on social media to our messaging service.

In the future, we want to increase the speed of these systems by ensuring they can get real-time labels of abusive accounts they have missed and good accounts they incorrectly identified as abusive. This will help our machine learning models adapt and respond immediately to any new kind of abuse. The more we can increase the cost for those who attempt to abuse our platform, the less they will succeed.

Addressing Challenges Beyond Our Product

We have also seen that some attackers attempt to modify WhatsApp software and trade unauthorized APK files to get around the constraints coded into the client itself. These modifications cannot circumvent our detection systems that run on our servers and apply to all users without exception. Still, we are improving our ability to detect these modified versions of WhatsApp as they both violate our Terms of Service and pose a security risk to users that have installed them on their device. Users should only download WhatsApp from our website or an App Store like Google Play or the Apple App Store.

We also go beyond the platform where possible. We work with providers to ban third-party apps that claim to provide unauthorized services for our platform. In addition, when we see companies claiming they can send bulk messages on WhatsApp, we demand they cease operations and reserve all rights available to us under applicable law.

Privacy and Safety

We strongly believe that providing a platform for private communication contributes to the safety of our users. People expect their messages to be secured so they do not fall into the wrong hands. In some places WhatsApp provides a lifeline to human rights organizations and journalists. Through a combination of efforts, we can help prevent abuse and continue to provide the privacy and security benefits that come with end-to-end encryption.

Fighting automated and bulk messaging is just one way that we help users stay safe. We also modify the way our product is designed based on how it is used. We now label forwarded messages to notify users when the message they received was not created by the sender. We have limited how many forwarded messages can be sent at once. In several countries we have launched public education campaigns to help users spot hoaxes and caution them to not spread rumors.

We take a broad view of our responsibility to assist users and this comprehensive approach helps maintain privacy and safety on WhatsApp. We encourage users who receive unwanted messages to block and report accounts so we can continue to improve WhatsApp and provide a valuable service around the world.

PX0356

We protected your account

If we notice any strange activity on your Pinterest account, we'll send you an email, reset your password, and log everyone out (including you).

Strange activity includes things like logins from unusual locations, many logins within a short period of time, or spammy behavior.

Get back on Pinterest

To get back on Pinterest, reset your password. Or, if your account's connected to Google or Facebook and you've logged in that way before, log in with Google or Facebook.

Once you're back into your account, you can add extra security to prevent this from happening again.

Add extra security to your account

- Turn on two-factor authentication
- Connect your account to Google or Facebook
- Make sure your email address is up to date
- Create a password for Pinterest that you don't use anywhere else

Other articles

Access your account +

Manage your account email +

Manage multiple accounts +

Set your preferences +

Connect your account +

Protect your account —

Review recent logins

Protect your account

Remove your email from someone else's account

Two-factor authentication

You got an email about a change you didn't make

You're seeing boards or Pins you didn't add

We protected your account

Review account block or suspension +

Still need help? Contact us

Was this article helpful?

Pinterest

Languages

English (US)

About us Our policies More info

What's Pinterest

Our Pinterest page

Engineering blog

Brand guidelines

Careers

Help Center

Pinterest Labs

Copyright

Personalized ads

Terms of service

Privacy

For businesses

For developers

For press

For investors

© Pinterest 2024

PX0359

About MaxMind

Founded in 2002, MaxMind is an industry-leading provider of IP intelligence and online fraud detection tools.

MaxMind provides IP intelligence through the GeoIP brand. Over 5,000 companies use GeoIP data to locate their Internet visitors and show them relevant content and ads, perform analytics, enforce digital rights, and efficiently route Internet traffic. Businesses can obtain additional insights into their customers' connection speeds, ISPs, and more using GeoIP data. MaxMind's industry-leading minFraud service helps businesses prevent fraudulent online transactions and reduce manual review. The minFraud service is used to screen over 175 million e-commerce transactions and account registrations a month. Over 7,000 e-commerce and other online businesses benefit from the minFraud service through our client and partner networks.

GeoIP2 Web Services

GeoIP2 Databases

GeoLite2 Free Geolocation Data

Affiliate Program

Your Privacy Choices

Notice of Collection

×

© 2024 MaxMind, Inc. MaxMind, GeoIP, minFraud, and related trademarks belong to MaxMind, Inc.

Terms of Use | Privacy Policy

PX0360

IP2LOCATION

IP Geolocation

Leveraging the power of **IP geolocation** and **Proxy detection** to refine your targeting strategies.

TRY DEMO VIEW PRICING

- Country, Region, City
- Latitude & Longitude, ZIP Code
- Area Code, Elevation
- ASN & District
- Weather Station Data

IP2LOCATION

- Address Type & Usage Type
- Time Zone, ISP, Domain, Net Speed
- Advertising Category
- VPN, Residential Proxy
- Mobile Carrier Information, MNC, MCC

Accurate IP Address Data

Unleash the power of accurate IP address data**IP2Location** is a non-intrusive IP location lookup technology that retrieves geolocation information with no explicit permission required from users.

Supports IPv4 & IPv6

It works for all IP addresses including IPv4 and IPv6 in one database or API. Simple and no extra cost required for IP location lookup.

Easy Integration

Integrate it effortlessly into your preferred software platforms to retrieve precise IP geolocation information using IP database, REST API and SDK (Java, .NET, PHP, Ruby, Python, Perl and more)

IP2LOCATION

Multiple Granularity

It comes with different IP database packages with varying levels of IP geolocation information granularity to suit your business needs. Pay for what you need.

Understanding Internet traffic via IP lookup

Experience the IP address lookup to access your visitors' detailed geolocation information

Your IP Address
Loading...

Country
Loading...

Region
Loading...

City
Loading...

Coordinates
Loading...

ISP
Loading...

Time Zone
Loading...

AS Number
Loading...

IP2LOCATION

Loading...

IDD & Area Code
Loading...

ZIP Code
Loading...

Weather Station
Loading...

Elevation
Loading...

Usage Type
Loading...

AS Name
Loading...

Is Proxy
Loading...

Proxy Type
Loading...

Proxy ASN
Loading...

Security Threat
Loading...

Proxy Last Seen
Loading...

Address Type
Loading...

Category
Loading...

District
Loading...

IP2LOCATION

Try IP2Location Demo

136.226.19.87 LOOK UP

Try out the demo to get the comprehensive geolocation data of an IP address. Supports both IPv4 and IPv6 address lookup.

Our Products

Discover our products for precise location and proxy detection

IP2Location IP GeoLocation

Determine your website visitors geographical location information and where they are coming from.

Accurate and Comprehensive IP geolocation and network data.

IP2Proxy Proxy Detection

Mitigate cybersecurity threats, safeguard against online fraud, and control access to your content and services effectively.

Detect anonymous proxy, VPN, TOR, SES, RES and data center IPs.

IP2LOCATION

Why IP2Location

37
Data Sets

2002
Established Since

20+
Programming Languages

20,000+
Developers

20B+
API Calls

24%
Fortune 500 Customers

800+
Research Citations

IP2LOCATION

Free Tools

Explore our free blog, SDK libraies and plugins to ease your works

Blog

Free articles and tutorials for guidance on geolocating user locations using databases and APIs.

Learn More →

Libraries

Free open-source libraries to simplify application integration through database usage.

Learn More →

Plugins

Free plugins for various platforms, such as blog, forum, CMS, E-commerce and so on.

Learn More →

IP2LOCATION

Unlock Geolocation Insights For Free

Experience accurate geolocation data with our complimentary IP2Location service.

TRY IT FOR FREE

Our Site

Client Portal
Online Demo
Solutions
Developers
Blog
Media Kit
Newsletter
Careers
Feature Request
Do Not Sell My Info
Contact

Free Tools

Applications
Firewall IP List
Search Engine IP List

Database

IP2Location™ Database
IP2Proxy™ Database

Web Service

IP2Locaton.io Web Service
IP2Location™ Batch Service
IP2Proxy™ Batch Service

Newsletter

Subscribe to our newsletter and stay up to date with the latest news and deals!

IP2LOCATION

Downloader Script

Extensions / Plugins

Widgets

IP Address Report

World Country List

See our 1,124 reviews on ★

© 2001 - 2024 IP2Location.com. All Rights Reserved. Terms of Service | Privacy Policy| Cookie Notice

PX0361

Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization

Jeremy Elson, John R. Douceur, Jon Howell
Microsoft Research
{jelson,johndo,howell}@microsoft.com

Jared Saul
Petfinder, Inc.
jared@petfinder.com

ABSTRACT

We present Asirra (Figure 1), a CAPTCHA that asks users to identify cats out of a set of 12 photographs of both cats and dogs. Asirra is easy for users; user studies indicate it can be solved by humans 99.6% of the time in under 30 seconds. Barring a major advance in machine vision, we expect computers will have no better than a 1/54,000 chance of solving it. Asirra's image database is provided by a novel, mutually beneficial partnership with Petfinder.com. In exchange for the use of their three million images, we display an "adopt me" link beneath each one, promoting Petfinder's primary mission of finding homes for homeless animals. We describe the design of Asirra, discuss threats to its security, and report early deployment experiences. We also describe two novel algorithms for amplifying the skill gap between humans and computers that can be used on many existing CAPTCHAs.

1. INTRODUCTION

Over the past few years, an increasing number of public web services have attempted to prevent exploitation by bots and automated scripts, by requiring a user to solve a Turing-test challenge (commonly known as a *CAPTCHA*[1] or *HIP*[2]) before using the service. Because the challenges must be easy to generate but difficult (for non-humans) to solve, all CAPTCHAs rely on some secret information that is known to the challenger but not to the agent being challenged. For our purposes, we can divide CAPTCHAs into two classes depending on the scope of this secret. In *Class I* CAPTCHAs, the secret is merely a random number, which is fed into a publicly known algorithm to yield a challenge, somewhat analogous to a public-key cryptosystem. *Class II* CAPTCHAs employ both a secret random input and a secret high-entropy database, somewhat analogous to a one-time-pad cryptosystem. A critical problem in building a Class II CAPTCHA is populating the database with a sufficiently large set of classified, high-entropy entries.

Class I CAPTCHAs have many virtues. They can be concisely described in a small amount of software code; they have no long-term secret that requires guarding; and they can generate a practically unbounded set of unique challenges. On the other hand, their most common realization—a challenge to recognize distorted text—evince a disturbingly narrow gap between human and non-human success rates. Optical character recognition algorithms are competitive with humans in recognizing distinct characters, which has led researchers toward increasing the difficulty of segmenting an image into distinct character regions [11]. However, this increase in difficulty affects humans as well. Although laboratory experiments suggest that humans can segment text characters accurately [2], CAPTCHAs deployed on commercial public web sites continue to use cleanly segmented challenges (e.g., Fig. 2a), some of which are dialed-down versions (Figs. 2d and 2e) of CAPTCHAs with difficult segmentation challenges (Fig. 2c). The owners of commercial web sites have apparently decided that a user's success at navigating a CAPTCHA depends not only on whether they are *able* to solve the challenge, but also on whether they are *willing* to put forth the effort. Informal discussions with MSN and other web site owners suggest even relatively simple challenges can drive away a substantial number of potential customers.

Class II CAPTCHAs have the potential to overcome the main weaknesses described above. Because they are not restricted to challenges that can be generated by a low-entropy algorithm, they

Figure 1: An Asirra challenge. The user selects each of the 12 images that depict cats. As the mouse is hovered over each thumbnail, a larger image and "Adopt me" link appear. "Adopt me" first invalidates the challenge, then takes the user to that animal's page on Petfinder.com.

[1] "Completely Automated Public Turing test to tell Computers and Humans Apart." CAPTCHA is a trademark of Carnegie Mellon University.
[2] "Human Interaction Proof"

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CCS'07, October 29–November 2, 2007, Alexandria, Virginia, USA.
Copyright 2007 ACM 978-1-59593-703-2/07/0011 ...$5.00.

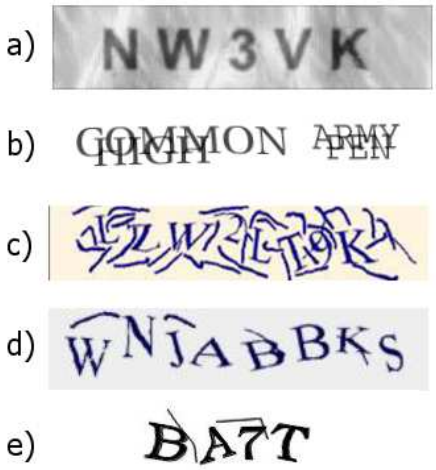


Figure 2: A gallery of text CAPTCHAs. Simple text challenges, such as *a* (register.com), are still common despite recent defeat by optical character recognition. Researchers have begun to focus on schemes that make letter segmentation difficult, as seen in *b* (Carnegie Mellon [13]) and *c* (Microsoft Research [2]). Webmasters, wary of what users will tolerate, dial back researchers' noise parameters, seen in *d* (Microsoft Hotmail) and *e* (Yahoo! Mail).

can exercise a much broader range of human ability, such as recognizing features of photographic images captured from the physical world. Such challenges evince a broad gulf between human and non-human success rates, not only because general machine vision is a much harder problem than text recognition, but also because image-based challenges can be made less bothersome to humans without drastically degrading their efficacy at blocking automatons.

A significant issue in building a Class II CAPTCHA is populating the secret database. Existing approaches take one of two directions: (a) mining a public database or (b) providing entertainment as an incentive for manual image categorization. Examples of the first group include the seminal work by Chew and Tygar [3], which used Google Image Search [6]; hotcaptcha [1], which references the HotOrNot database [7]; and KittenAuth [16], which draws images from Wikimedia Commons [17]. A problem with these approaches is that the public source of categorized images is small or available to attackers, so a small, fixed amount of effort spent reconstructing the private database can return the ability to solve an unbounded number of challenges. The second direction was pioneered by the ESP-PIX CAPTCHA [13], whose database is populated as a deliberate side effect of playing the ESP Game [15], a very clever mechanism for enticing people to label images accurately. Although potentially powerful, it is not yet clear whether this approach will yield a sufficiently large set of categorized images. Furthermore, many of the images in the current implementation are rather abstract, which may make the challenge difficult enough to drive away users.

In this paper, we present a new direction for populating image databases for Class II CAPTCHAs, namely re-purposing a large, continually evolving, private database of images that are manually categorized. Although this may seem trivial, it is not *a priori* clear why the owner of such a database would be willing to release the images for use in Turing-test challenges. The answer is that there can exist—and, in at least one instance, does exist—an *alignment of interests* between a database owner and web-service owners wishing to secure their sites. Both parties can benefit from selective, wide-scale display of categorized images: the latter for security and the former for advertising.

We present **Asirra**[3], a CAPTCHA that asks users to categorize photographs depicting either cats or dogs. An example is shown in Figure 1. Asirra's strength comes from an innovative partnership with Petfinder.com [9], the world's largest web site devoted to finding homes for homeless animals. Petfinder has a database of over three million cat and dog images, each of which is categorized with very high accuracy by human volunteers working in thousands of animal shelters throughout the United States and Canada. Petfinder has granted ongoing access to its database, which grows by nearly 10,000 images daily, to the Asirra project. In exchange, Asirra provides a small "Adopt me" link beneath each photo, promoting Petfinder's primary mission of exposing adoptable pets to potential new owners. This partnership is mutually beneficial, and also produces the dual social benefits of improving computer security and animal welfare.

This paper describes Asirra and an analysis of its strengths and weaknesses. We also report our deployment experience, and the results of two user studies involving 332 test subjects.

Asirra is easy for users; it can be solved by humans 99.6% of the time in under 30 seconds (Section 6, Table 1). Barring a major advance in machine vision or compromise of our database, we expect computers will have no better than a 1/54,000 chance of solving it (Section 6, Table 2). Anecdotally, users seem to find the experience of using Asirra much more enjoyable than a text-based CAPTCHA that provides equal security.

The organization of this paper is as follows. In Section 2, we review related work in more detail. In Section 3, we describe the design of Asirra. §3.1 describes user experiments we performed to quantify humans' performance. §3.2 explores the other side of the equation—potential attacks on Asirra, and how they can be resisted. We developed two algorithms that can be used to improve virtually all CAPTCHAs, including those that are text-based; these improvements are described in Section 4. In Section 5 we describe our scalable Asirra implementation. Finally, in Section 6, we summarize our contributions and offer conclusions.

Asirra is available free at `www.asirra.com`.

2. RELATED WORK

Since the concept of a CAPTCHA was widely introduced by von Ahn in 2000 [14], hundreds of design variations have appeared. By far, most are text-based: The computer generates a challenge by selecting a sequence of letters, rendering them, distorting the image, and adding noise. Text CAPTCHAs are popular because they are simple, small, and easy to design and implement. Challenges as short as four characters are robust against random guessing; there are $36^4 \approx 1.7$ million possible four-character challenges consisting of case-insensitive letters and digits.

Unfortunately, computers can do far better than guess randomly. Simard et al. showed that Optical Character Recognition (OCR) can achieve human-like accuracy, even when letters are distorted, as long as the image can be reliably segmented into its constituent letters [11]. Mori and Malik demonstrated that von Ahn's original GIMPY CAPTCHA [13] can be solved automatically 92% of the time [8].

Consequently, recent text-based CAPTCHAs have focused on making image segmentation difficult. Figure 2c shows a challenge designed by Chellapilla et al., who claim it is hard for OCR because the noise confounds known segmentation techniques [2]. Microsoft's Hotmail (free email service) deployed it; however, due

[3]"Animal Species Image Recognition for Restricting Access"

to usability concerns, they later selected noise parameters demonstrated in Figure 2d. Yahoo's current CAPTCHA, shown in Figure 2e, seems to have suffered a similar fate. The noise is not sufficient to make automatic segmentation unreliable. Text CAPTCHAs seem either too easy to be secure, or too difficult to be tolerated by users.

2.1 Image Classification CAPTCHAs

Text-based CAPTCHAs seem to universally suffer from an unfortunate property: Making them hard for computers also makes them hard for humans. This has led some researchers to use photographs as CAPTCHAs instead. Because general machine vision is a much harder problem than character recognition, there are opportunities to find and exploit larger gaps in the capabilities of humans and computers.

Chew and Tygar [3] were among the first to describe using labelled photographs to generate a CAPTCHA. They generated a database of labelled images by feeding a list of easily-illustrated words to Google Image Search [6]. Unfortunately, this technique does not yield well-classified results due to Google's method of inferring photo contents based on surrounding descriptive text. To use Chew and Tygar's example, the word *pumpkin* may refer to either a large vegetable or someone's pet cat Pumpkin. Because of these errors, they manually cull bad images from their collection. This is devastating to the security of the scheme. A database small enough to be manually constructed by researchers is also small enough to be manually *re*-constructed by an attacker.

Of course, applying automation to database construction is inherently problematic. If a researcher can populate a database by writing a program to automatically classify images, an attacker can write a program to beat the CAPTCHA by performing the same classification.

A novel solution to this problem is described by von Ahn *et al.*: They were able to entice humans to manually describe images by framing the task as a game. Their "ESP Game" awards points to teams of non-communicating players who can both pick the same label for a random image, encouraging them to use the most obvious label [15]. Their PIX CAPTCHA displays four images from the ESP Game database that have the same label, then challenges the user to guess the label from a menu of 70 possibilities.

PIX is clever, but has several potential problems. First, its scale seems insufficient. By solving PIX repeatedly, it is not hard to get repeated images, making the database easy to reconstruct by an attacker. However, perhaps more fundamental, it has a fixed menu of only 70 object classes. This makes it a potential target for brute force attacks (though potentially defensible using our token bucket scheme; see Section 4.2). Even with a large number of categorized images, it may be difficult to add a large number of classes. As the number of classes goes up, so does the number of words that could reasonably be used to describe a set of photos. Finally, PIX photos are sometimes abstract, making it potentially difficult or frustrating as a CAPTCHA.

A fascinating use of a large-scale human-generated database is the site HotCaptcha.com. HotCaptcha displays nine photographs of people and asks users to select the three which are "hot." Its database comes from HotOrNot.com, a popular web site that invites users to post photos of themselves and rate others' photos as "hot" or "not." HotCaptcha is clever in its use of a pre-existing motivation for humans to classify photos at a large scale. However, humans may have difficulty solving it because the answers are subjective and culturally relative; beauty has no ground truth. It is also offensive to many people, making it difficult for serious web sites to deploy.

Finally, worthy of mention is the similar-seeming KittenAuth project [16]. Like Asirra, KittenAuth authenticates users by asking them to identify photos of kittens. However, this is a coincidental and superficial similarity. KittenAuth is trivial to defeat because it is has a database of less than 100 manually selected kitten photos. An attacker can (indeed, already has [12]) expose the database by manually solving the KittenAuth challenge a few dozen times. An arbitrary number of challenges can then be solved using an image comparator robust to simple image distortions.

3. ASIRRA

Asirra surmounts the image-generation problem in a novel way: by forming a partnership with Petfinder.com [9], the world's largest web site devoted to finding homes for homeless pets. Asirra generates challenges by displaying 12 images from a database of over three million photographs that have been manually classified as cats or dogs. Nearly 10,000 more are added every day by volunteers at animal shelters throughout the United States and Canada. The size and accuracy of this database is fundamental to the security provided by Asirra; it is what differentiates our work from previously proposed image-based CAPTCHAs.

In exchange for access to Petfinder's database, Asirra provides an unobtrusive "Adopt me" link beneath each photo. This promotes Petfinder's primary mission of exposing adoptable pets to potential new owners. To maximize the probability of successful adoptions, Asirra will employ IP geolocation to determine the user's approximate region, and preferentially displays pets that are nearby. The security implications of this feature are discussed in Section 3.2.1.

Asirra has several attractive features:

- Humans can solve it quickly (§3.1.2) and accurately (§3.1.3).
- Computers can not solve it easily (§3.2).
- Unlike many image-based CAPTCHAs which are abstract or subjective, Asirra's challenges are concrete, inoffensive (cute, by some accounts), require no specialized or culturally biased knowledge, and have definite ground truth. This makes Asirra less frustrating for humans. Some beta-testers found it fun. The four-year-old child of one asked several times to "play the cat and dog game again."
- It promotes an additional social benefit: finding homes for homeless animals.

Asirra also has several disadvantages:

- Most CAPTCHAs are implemented as stand-alone program libraries that can be integrated into a web site without introducing external dependencies. In contrast, Asirra, like PIX, is both an algorithm and a *database*; there is only one instance of it. Therefore, Asirra must be implemented as an administratively centralized web service that is used to generate and verify CAPTCHAs on-demand for every site that wishes to use it. (Our scalable implementation is described in Section 5.)
- Asirra may abruptly lose its security if the database is compromised. For example, an attacker may hire cheap labor to classify all three million images. If this does happen, we may not even be aware of the attack.
- A typical Asirra challenge requires more screen space than a traditional text-based CAPTCHA.
- Like virtually all other CAPTCHAs, Asirra is not accessible to those with visual impairments (§3.3).

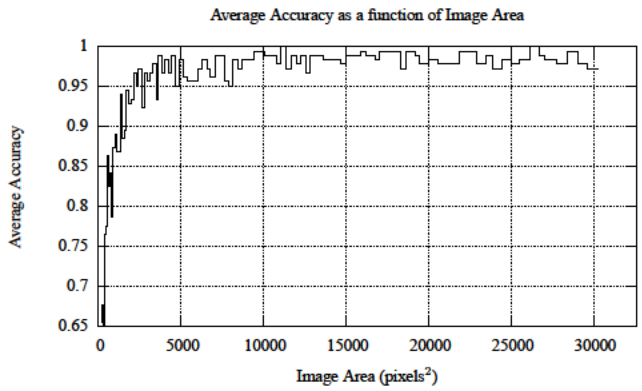


Figure 3: Size of an individual image vs. classification accuracy in our test population.

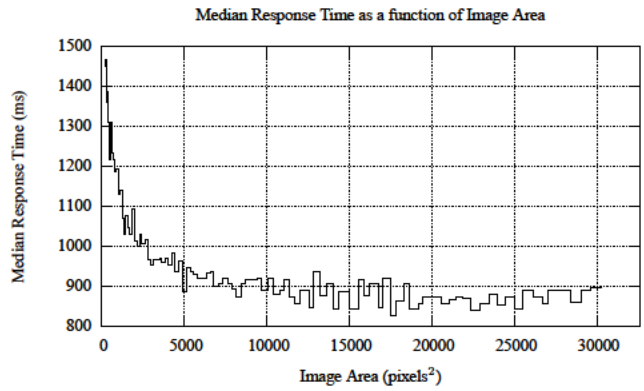


Figure 4: Size of an individual image vs. time required to classify it in our test population.

3.1 Usability

To quantitatively evaluate Asirra, we performed several user studies. We also collected data from our live Asirra deployment, which is described in Section 5.

The user studies, in total, displayed 23,208 cat and dog images to 332 test subjects. The subjects were Microsoft employees, friends, and family, recruited via postings to internal mailing lists. Our experiment displayed a random sequence of cat and dog images, one at a time. Users were asked to identify the species depicted in each image as it appeared. We used 300 random images from the Petfinder database, each scaled to a random size as described in the next section. The users' response time and accuracy were recorded. To match the conditions of a real CAPTCHA, we implemented the experiment in the web browser using HTML and JavaScript, and asked users to participate from their own home or office computers.

The real Asirra web service went live in March of 2007. In the subsequent 6 weeks, it served about 100,000 challenges. Most of these were from our own web page that demonstrates Asirra. However, several dozen web sites (blogs, free services, etc.) have experimented with integrating Asirra and were responsible for about 13,000 "real" challenges.

In following sections, we characterize various aspects of Asirra's performance, based on both the outcome of our user experiments and data from our deployment.

3.1.1 *Best Image Size*

Our first design question was, "What is the best image size to display?" We expected larger images to exhibit higher accuracy and faster response time. However, smaller images have the advantage of taking up less screen space, making Asirra easier to integrate visually with the rest of a web page. Smaller images are also faster to download, which especially important for users with slow Internet connections.

To find the best image size, our first user experiment randomly varied the size of the images from a minimum of 225 total pixels (about 15x15 pixels square) to a maximum of 30,000 pixels (about 175x175 pixels square). We used total pixels instead of linear dimension as our metric because most images are not square. We collected data from 18,311 displays of images to 185 users.

Figure 3 plots image size vs. average accuracy. Each graph segment depicts the average of 1/100th (183) of our data points. 10,000 pixels seems to be the sweet spot: larger images show no improvement. (We quantify classification accuracy in Section 3.1.3).

3.1.2 *Typical Response Time*

We also used our first experiment to evaluate the effect of image size on response time. Figure 4 shows the results. We plot the median response time because it is robust against outliers (e.g., when a subject receives a phone call during the experiment). 10,000 pixel images are typically classified in about 900msec. There does seem to be a slight (50msec) speed benefit to displaying images larger than 10,000 pixels. It is also interesting to note from Figures 3 and 4 that smaller images cause degradation in both response time *and* accuracy. That is, users spend longer looking at an image but still end up getting it wrong.

Budgeting 900msec per image, plus a few extra seconds to understand the task, we expect most users will spend about 15 seconds solving a single, 12-image Asirra challenge.

3.1.3 *Classification Accuracy*

After our first experiment made the best image size clear, we ran a second user experiment focusing exclusively on images scaled to 10,000 pixels. Our goal was to collect a large number of samples at our target image size. We collected data from 4,717 displays of images to 147 users. The overall accuracy rate was 98.5%.

The Asirra challenge has 12 images. Based on a 98.5% per-image accuracy, 83.4% of users should be able to pass Asirra after one challenge, 97.2% after two challenges, and 99.5% after three. However, Asirra uses a novel scheme called the *Partial Credit Algorithm*, described in Section 4.1, which significantly increases the probability of users passing Asirra while only marginally improving the yield of a bot. With PCA, we expect 99.6% of users will solve Asirra after two challenges.

Extracting a good estimate of users' error rates in the wild is somewhat more difficult. In our controlled experiment, users were focused exclusively on image categorization, so errors were clearly attributable to task difficulty. In contrast, the deployed CAPTCHA is integrated into web forms with many other fields and information. Any click of the form's "Submit" button causes Asirra to score the challenge, even if the user had some other intent in mind; 26% of scoring requests received by our server had no response at all (no cats selected). Of non-null responses scored, 66% were scored as correct. Note that in this accounting, a user who successfully solves Asirra on her second try counts as 50% accuracy (one right plus one wrong response). In addition, since Asirra is new, users may sometimes be tinkering: "Oh, what happens if I get one wrong?"

Figure 5: Though rare, some images at Petfinder.com are confusing to humans. *(left)* A photo depicting a cat and dog together. *(right)* A dog-like cat.

3.1.4 Automatic Database Pruning

The classification errors made by users are not uniformly distributed: some images are confusing, even to humans. Figure 5 shows two examples.

If Asirra achieves only modest popularity, our accuracy figures are likely to stay close to those described in the previous section. However, if Asirra receives widespread use (tens of thousands of challenges generated per day), we will have enough data to make automatic, statistically significant judgements about confusing images. There will be cases when a user ultimately succeeds in passing Asirra, but may have made errors along the way. Once Asirra decides a user is human, it will be able to mark the previously-incorrect images as "possibly confusing." Images that are marked beyond a threshold can then be removed from the database.

3.1.5 Interest in the "adopt me" link

In both user experiments, we placed a small "Adopt me" link below the cat or dog displayed in each trial. We did not give users any instructions regarding this link or tell them what it did. Our goal was to test how often Internet users involved in some unrelated task would be motivated (by curiosity, cuteness of a photo, etc.) to click it. This test is important because it suggests whether our partnership with Petfinder is viable.

In our controlled experiments, 27 users (7.5% of the test population) clicked "Adopt me". One of our beta-testers adopted a beagle.

The public's interest in the link has been lower. In the 6 weeks following Asirra's release, we issued 13,334 "real" challenges—that is, integrated with pages other than our own demonstration pages. 279, or 2.1%, led to clicks on "Adopt me".

We discuss the security implications of "Adopt me" in §3.2.1.

3.2 Security

Before discussing the security of Asirra, it is useful to review the threat model. CAPTCHAs are an unusual area of security in that we are not trying to provide absolute guarantees, only slow down attackers. By definition, anyone can "break" a CAPTCHA by devoting a small amount of human effort to it. A CAPTCHA therefore is successful, essentially, if it forces an automated attack to cost more than 30 seconds worth of a human's time in a part of the world where labor is cheap. The generally accepted figure in the literature [13, 2] seems to be that CAPTCHAs should admit bots less than with less than 1/10,000 probability.

3.2.1 Attacks on "Adopt Me"

The most common first question we get when people see Asirra is, "Doesn't the adopt-me link defeat your security?" This misconception is understandable; each pet's page on Petfinder.com describes it as being either a cat or dog. However, this is not actually a security hole. The adopt-me links are not direct links to Petfinder.com; they lead to the Asirra web service. Asirra provides a redirection to Petfinder.com only after it has marked the challenge as invalid. (A new challenge is then fetched and displayed.) Asirra rejects attempts to solve invalidated challenges. In addition, Asirra only permits redirection for a single adopt-me link per 12-image challenge. The number of allowed redirections per IP address per day is also limited, to prevent adopt-me from becoming a vector for revealing large portions of the database.

Adoption links also have a second, more subtle effect on security. Currently, the images shown as a challenge are selected at random from the entire database of pet images. To maximize its utility to Petfinder, we would ultimately like to restrict the challenge to pets that are close to the user based on IP geolocation, and currently available for adoption. These constraints reduce the usable database for a given user to just a few thousand images, small enough to be easily exploitable. Our plan, therefore, is to allow the first few challenges per day from an IP address to use the restricted database; subsequent challenges will be drawn from the complete collection. We have not yet implemented this feature, so our challenges currently draw images from Petfinder's complete image pool, including the history of pets previously available for adoption.

3.2.2 Brute Force

The simplest attack on Asirra is brute force: Give a random solution to challenges until one succeeds. If an attacker has no basis for a decision (that is, a 50% probability of success for each image) brute force will succeed with probability 1/4,096 for a 12-image challenge. This is a large enough slowdown that it becomes easy to detect and evade such an attack. We use a token bucket scheme, described in detail in Section 4.2. Briefly, the scheme penalizes IP addresses that get many successive wrong answers by forcing them to answer two challenges correctly within 3 attempts before gaining a ticket. With this scheme in place, attackers can expect only one service ticket per 5.6 million guesses.

3.2.3 Machine Vision Attacks

While random guessing is the easiest form of attack, various forms of image recognition can allow an attacker to make guesses that are better than random. Asirra's strength, however, comes not only in the size of its image database, but its diversity. Photos have a wide variety of backgrounds, angles, poses, lighting, and so forth—factors that make accurate automatic classification difficult. As Figure 6 demonstrates, the variations between photos are large, but visual differences between cats and dogs are often subtle.

Based on a survey of machine vision literature and vision experts at Microsoft Research, we believe classification accuracy of better than 60% will be difficult without a significant advance in the state of the art. For example, the best known algorithms for facial recognition can achieve in excess of 90% accuracy under controlled conditions, but are not robust to occlusions and variations in pose similar to those seen in our database [5, 18]. The 2006 PASCAL Visual Object Classes Challenge [4] included a competition to identify photos as containing several classes of objects, two of which were Cat and Dog. Although cats and dogs were easily distinguishable from other classes (e.g., "bicycle"), they were frequently confused with each other.

One attack on our database proposed by vision researchers was color histogram analysis. We implemented and tested this technique. We first characterized each of 150 random images as a 15-feature color vector, computing histograms of each fifth of the red,

Figure 6: The differences between cats and dogs are immediately obvious to humans. In many cases, species look similar, with only subtle cues to distinguish them. This makes it a hard vision problem.

green, and blue channels of each image. We then computed the best least-squares fit coefficients of these 15 features to the two (cat and dog) image classes. The coefficients were then used to predict the class of another 150 images. The result was 56.9% accuracy.

If we assume an attacker can build a 60% accurate classifier, the probability of solving a 12-image challenge improves from 1/4,096 to 1/459. However, with our token bucket scheme in place (§4.2), this is reduced to one ticket per about 70,000 guesses.

3.2.4 Database Attacks

One way to attack any image-based CAPTCHA is to reconstruct its underlying database. In a sense, this attack is unavoidable: A portion of the database is revealed each time a challenge is displayed. By solving enough challenges manually, the entire database is eventually revealed to the attacker. The key question is not "is database reconstruction possible?" but rather "is database reconstruction economical?"

A CAPTCHA can be considered secure if the most cost effective way to defeat it is for a human to solve it once per unit of service gained. In other words, our goal is to make the most efficient attack an on-demand one (e.g., pay people to sign up for one Hotmail account at a time). This tradeoff is directly informed by the size and stability of the underlying database.

Oli Warner's KittenAuth provides an instructive example. It originally launched with only 42 images; within days, someone posted a map of those images' MD5 hashes to a cat-or-dog bit [12]. Classification of 42 images does not take much investment, so this form of attack was an easy choice. Asirra's three million images moves the break-even point significantly. An attacker would expect to solve about 750,000 12-image challenges to reveal 95% of the database. Reconstruction of our database is not economical unless there is a financial incentive to solve millions of Asirra challenges.

Attackers might also try to reconstruct Asirra's database by attacking its source—that is, automating a long series of queries to Petfinder.com. However, Petfinder's public interface only displays pets currently available for adoption, which represents less than 10% of their total historical database. In addition, there is no efficient way to track database changes using Petfinder's public interface. The private API provided to Asirra by Petfinder is not available to the public.

Note that in this section we do not consider techniques for reusing an individual image such that it appears to be multiple distinct images. There are robust image comparator functions (e.g., image hashes, color histograms) that are insensitive to many simple image distortions. Warping an image sufficiently to fool a computer will likely also be troublesome to a human—the arms race found in text CAPTCHAs that we are trying to avoid.

3.2.5 Implementation Attacks

Many CAPTCHAs suffer from weak implementations. For example, some allow a session ID authorized by a single successful challenge to be re-used repeatedly to gain access to a protected service. Some assume trustworthy clients.

To avoid such weaknesses, we have taken two approaches. First, we made the Asirra web service as conceptually simple as possible, so that we can verify its correctness via code review. (The web service is about 700 lines of Python code.) Second, as we discuss further in Section 5.1, we designed the Asirra interface so as to minimize the implementation burden on the webmaster. For example, we do not require sites that use Asirra to keep track of user sessions or any other form of state. Our example service is, in fact, completely stateless.

3.3 Accessibility

Virtually all visual CAPTCHAs, including Asirra, are not accessible to visually impaired users. The problem of CAPTCHA accessibility is an important one, but beyond the scope of this paper. Accessible web sites typically allow users to switch between a visual challenge (e.g., distorted letters) and an auditory one (e.g., numbers being recited over noise). Asirra is only meant to replace visual CAPTCHAs. The audio-accessible version of a site's CAPTCHA is valuable, but orthogonal to Asirra.

However, it is worthwhile to note that Asirra's core idea of interest alignment is potentially usable in an audio CAPTCHA. For example, online music stores typically provide free samples of millions of songs in their catalogs. These might be used with a "buy me" link.

4. IMPROVEMENTS TO ALL CAPTCHAS

In the course of our work on Asirra, we developed two algorithms that can be used to improve virtually all CAPTCHAs, including those that are text-based. These techniques share a common theme. Imagine that users' responses to Asirra challenges were scored manually, by a human judge, instead of automatically by computer. Even in this seemingly straightforward task, the flexibility of human judgement would be a valuable asset. For example, our human judge could see the same user try to solve three challenges, getting 11 out of 12 images correct each time. She might say, "That looks like a human who is just having a little trouble. I'll let him pass." Conversely, our judge might see the same IP address submitting thousands of random, incorrect answers. Even when finally given a correct answer, she'd still suspect the user was a bot.

In the next sections, we describe two algorithms designed to make CAPTCHA scoring a little more "human": the Partial Credit Algorithm (§4.1) and CAPTCHA Token Buckets (§4.2).

4.1 The Partial Credit Algorithm

Traditionally, computers score CAPTCHA responses with one bit of output: right or wrong. The intuition behind the Partial Credit Algorithm (PCA) is that *a user's response contains much more than one bit of information*. Two "almost right" answers are strong evidence that a user is human; two random answers are not. Without PCA, this distinction is lost.

The idea of an "almost right" response is meaningful in the context of a wide variety of CAPTCHAs. For example, in a text-based CAPTCHA, typing five out of six characters correctly can be considered almost-right. In Asirra, we consider a response to be almost-right if 11 out of the 12 images presented are identified correctly.

Asirra awards Partial Credit to a user who gets a challenge almost right. If the user gets a subsequent challenge almost right—that is, while already holding Partial Credit—we judge the response as if it were completely correct. (Asirra challenges all take place within a *session*, making it easy to associate a single user's responses.)

The appeal of PCA is that its use of previously-ignored information allows us to significantly improve the pass rates for humans while minimally improving the pass rates for bots. For example, in Asirra, PCA reduces the number of humans rejected after 2 challenges from 2.8% to 0.4%: a 7-fold reduction in unhappy users (Section 6, Table 1). However, the bot yield improves from 1/4,096 to 1/3,952: only a 3% improvement (Section 6, Table 2).

In comparison, simply scoring every almost-right answer as correct (i.e., passing users who get 11/12) has a devastating effect on security: a random-guess bot's success rate improves from 1/4,096 to 1/315, a 13-fold increase.

We can mathematically model the effect of PCA as follows. We consider a new user to arrive in the unverified (u) state, and by solving a challenge the user moves into the verified (v) state. PCA introduces an intermediate (i) state, which the user moves into when getting a challenge almost right. From the intermediate state, if the user almost—or completely—solves a subsequent challenge, the user moves to the verified state; otherwise, the user is returned to the unverified state.

Let α represent the probability of solving a challenge, β be the probability of getting close enough to enter the intermediate state, and γ be the probability of getting close enough to become verified when in the intermediate state. After n steps, the probability that the user is in each state is given by the following recurrence relation:

$$\begin{aligned} u_n &= (1-\alpha-\beta)\,u_{n-1} + (1-\gamma)\,i_{n-1}, & u_0 &= 1 \\ i_n &= \beta\,u_{n-1}, & i_0 &= 0 \\ v_n &= v_{n-1} + \alpha\,u_{n-1} + \gamma\,i_{n-1}, & v_0 &= 0 \end{aligned}$$

The expected number of trials until verification is:

$$E = \sum_{1 \le n < \infty} n\,(v_n - v_{n-1}) = \frac{1+\beta}{\alpha + \beta\,\gamma}$$

In Asirra, the user moves to the intermediate state if exactly one image (out of 12) is misclassified; from the intermediate state, the user moves to the verified state if zero or one image is misclassified. Thus, using the binomial distribution function b, the PCA probabilities are:

$$\begin{aligned} \alpha &= b(0; 12, 1-p) \\ \beta &= b(1; 12, 1-p) \\ \gamma &= b(0; 12, 1-p) + b(1; 12, 1-p) \end{aligned}$$

This model can be used to compute the effect of PCA on any CAPTCHA that can define a solution as "almost right." The effect on 12-image Asirra challenges is shown in Tables 1 and 2.

4.2 CAPTCHA Token Buckets

Asirra, like virtually every CAPTCHA, is subject to brute-force attacks. Since guessing is nearly free, a success probability of 1/4,096 may not be a strong enough assurance that attackers can't automatically collect service tickets en masse.

Our countermeasure is based on the supposition that a bot can be identified as a small number of IP addresses that submit a very large number of incorrect responses, interspersed with a much smaller number of correct responses. (Assume for the moment that each user has a unique IP address; we will relax this constraint shortly.)

The scheme works by assigning a token bucket to every IP address that requests an Asirra challenge. Each bucket is initialized with *TB-Max* tokens. Every time a client submits *any* response to a challenge, a single token is removed from the client IP's bucket, down to a minimum of 0. Every time a client submits a *correct* response, *TB-Refill* tokens are added, up to a maximum of *TB-Max*. In our implementation, *TB-Max* is 100 and *TB-Refill* is 3. (The bucket is also re-initialized to *TB-Max* tokens after 24 hours of disuse, but this is not strictly required.) If a user submits a response while its token bucket is empty, the user is told the answer is incorrect, regardless of their response.

A brute-force bot trying to earn a large number of service tickets from Asirra will quickly empty its token bucket, as its incorrect guesses will outnumber its correct guesses by more than a factor of *TB-Refill*. In this steady-state, the empty-bucket policy will force

bots to correctly answer *two* challenges within *TB-Refill* attempts before they earn credit for a correct response.

Intuitively, the effect of this policy is to amplify the skill difference between humans and bots. Modelling it mathematically is straightforward. A bot has some probability p of getting a single challenge correct. With an empty token bucket, the bot must also get a second challenge right within *TB-Refill* tries; this happens with probability $1-(1-p)^{TB-Refill}$. That is, token buckets reduce the bot success probability to

$$p \times \left(1-(1-p)^{TB-Refill}\right)$$

For example, if a bot has a 1/4,096 chance of getting a single challenge correct, and *TB-Refill* is 3, our scheme reduces the probability of getting a ticket to one per about 5.6 million attempts per bot IP address.

4.2.1 Preventing denial-of-service against humans

Our scheme, as described so far, makes the assumption that each user has a unique IP address. That assumption is both strong and incorrect. Users share IP addresses when they access the Internet via a web proxy or firewall that performs network address translation. Such configurations are common in large corporations and some Internet Service Providers.

This is a problem for the token bucket scheme we described previously. When a human shares an IP address—and, thus, a token bucket—with a bot, the human is denied service while the bot is attacking. Since bots are much faster than humans, the human is likely to encounter an empty token bucket every time she submits an answer, and thus will always be marked wrong.

We address this problem by defining a slightly more complex scheme that uses two token buckets. The first is a per-IP bucket, as described before. The second is a per-*session* bucket. (All interactions with Asirra must take place within a session, created when a client first begins to interact with Asirra's web service.) The rules governing the buckets are:

- Per-IP buckets are given an initial value of *TB-Max* tokens, as before.
- When a new session is created, the per-session bucket is given an initial value equal to the current value of the client's per-IP bucket. Creation of a new session also deducts a token from the per-IP bucket.
- When any response is submitted, a token is deducted from *both* the per-IP and per-session bucket.
- When a correct response is submitted, *TB-Refill* tokens are added to *both* the per-IP and per-session bucket.
- A user is only told their response is correct if their *session* bucket is non-empty.

Intuitively, this scheme blacklists IP addresses, but allows humans to restore service to individual sessions by solving one extra challenge correctly.

Using a single, per-session bucket would not be effective because clients are untrusted. A bot does not play by the rules, and would simply create a new session before every guess. Therefore, the per-IP bucket is still necessary. However, legitimate users *will* create a single session and use it persistently, allowing us to distinguish them from bots sharing their IP address. A user sharing an IP with an attacking bot will initially get an empty per-session bucket, but will add *TB-Refill* tokens after her first correct response. These per-session tokens are not consumed by the bot, so the user is granted access after a second correct response.

5. DEPLOYMENT

Most CAPTCHAs are *algorithms*; their realization is typically source code that can be shared freely and integrated into web sites without external dependencies. In contrast, Asirra is both an algorithm and a *database* that must be kept secret. To protect the database, we had no option but to implement Asirra as a web service. In this section, we briefly describe our service's interface, implementation, and deployment.

The service is divided into the trusted web service, run by us, and the untrusted client component, implemented by us but run in the end-user's browser.

5.1 Interface

Our goal in designing the interface to the web service was to make it as easy as possible for the webmaster. Overly complex interfaces tend to be mis-used in ways that compromise security.

Suppose a web site wishes to use Asirra to protect a sign-up form. Our site has instructions for the webmaster describing how to include a few lines of JavaScript in the HTML form she wishes to protect. This code loads the bulk of Asirra's front-end implementation from the Asirra server.

Once the end-user views, completes, and submits the form—including solving the Asirra CAPTCHA—our JavaScript will set a hidden input field in the HTML form called `Asirra_Ticket`. The ticket is a bit string that the webmaster then must validate against the Asirra web service to ensure the client did not forge or re-use it. This validation typically happens as a step of processing the client's form on the webmaster's back-end server.

Ticket validation is simple: the ticket is passed to an Asirra validation URL. Our web service returns "Pass" only if the ticket is valid, recent, and has never been validated before. The JavaScript that manages user interactions on the client prevents the form from being submitted until a valid ticket has been received by passing the Asirra challenge. Thus, webmasters typically only see validations fail when their service is under attack.

5.2 Implementation

Asirra manages its own interaction with the user inside an embedded DIV. The interactions between client and web service are performed using "AJAX" (asynchronous HTTP requests executed from within JavaScript). When our JavaScript is loaded, it first creates the visual elements of the challenge in the user's browser. It then sends requests to the Asirra web service to create a new Asirra session, retrieve one or more challenges, and submit user responses for scoring. The client earns a service ticket for a correct response (or, in some cases, an almost-correct response; see Section 4.1). The user can choose to see a new challenge if they are asked to solve one that contains a confusing image.

Though administratively centralized, the web service's implementation is distributed to ensure scalability. Load balancing is achieved by mapping our web service's DNS name to multiple IP addresses; clients choose among them randomly. The first action performed by a client is session creation. Whichever machine is randomly selected by the client to execute this action becomes the permanent custodian of that session's state. The custodian stores the session state locally and returns a session ID to the client. The session ID has the custodian server's ID embedded in it.

If later operations that are part of the same session arrive at a different Asirra service machine, the request is forwarded to the session custodian by finding its identity embedded in the session ID. (The client is not trusted; a forged session ID will simply fail to find any expected state on the incorrectly indicated custodian server.) Our forwarding scheme ensures that at most two machines

Challenges Solved	Users Passed, No PCA	Users Passed, With PCA
1 (≈15 sec)	83.4%	83.4%
2 (≈30 sec)	97.2%	99.6%
3 (≈45 sec)	99.5%	99.96%

Table 1: Expected user population that will pass Asirra after 1, 2, and 3 challenges, with and without PCA (§4.1). Assumes a 12-image challenge, 98.5% classification accuracy for an individual image (§3.1.3), and 15 seconds required per challenge (§3.1.2).

Image Classifier Accuracy	Bot Success Rate		
	No PCA	With PCA	PCA + Tokens
50% (random guessing)	1/4,096	1/3,952	1/5.2 million
60% (known techniques)	1/459	1/404	1/54 thousand
70% (major AI advance)	1/72	1/54	1/990

Table 2: Expected success rates of bots with three hypothetical image classifiers. Assumes a 12-image challenge. The effects of PCA (§4.1) and Token Buckets (§4.2) with *TB-Refill* of 3 are also considered. A detailed threat analysis can be found in §3.2.

are ever involved in servicing a single request: the machine which receives the request from the client, and the machine that owns the session state and receives the sub-request. The Asirra service is therefore readily scalable; the overhead of parallelization will never be more than 2x regardless of the total size of the farm.

In practice, we have observed lower overhead; request forwarding is not the common case. Many combinations of browser and operating system continue to use the same IP address for a given domain name for the duration of a browser session. This may be an aspect of many browsers' "DNS Pinning" policy meant to reduce cross-site scripting attacks.

The web service is about 700 lines of Python code. It is currently deployed on Amazon's EC2 [10] compute cloud platform, making it easy to add resources as the service's popularity increases.

6. SUMMARY AND CONCLUSIONS

In this paper, we presented Asirra, a CAPTCHA that asks users to identify cats out of a set of 12 photographs of both cats and dogs. Our image database is provided by a novel, mutually beneficial partnership with Petfinder.com, which has a database of over three million images of pets looking for new homes. In exchange for the use of these images, Asirra displays an "adopt me" link beneath each one, promoting Petfinder's primary mission of finding homes for homeless animals. (Security implications of Adopt-Me are discussed in Section 3.2.1.)

Asirra is attractive because humans can solve it quickly and accurately (Table 1), but it provides significant protection from bots (Table 2). Asirra provides a social benefit: improving animal welfare. And, anecdotally, users report that Asirra is enjoyable: Spending a few moments looking at cute kittens is generally preferred to squinting at deformed letters.

Asirra is a free web service, available at `www.asirra.com`.

Our contributions in this paper also include two techniques that we use in Asirra, but can be added to virtually any CAPTCHA. The Partial Credit Algorithm (§4.1) significantly improves the pass rates of humans while only marginally improving the yield of bots. Conversely, our use of CAPTCHA Token Buckets (§4.2) significantly decreases the yield of brute-force bots while having a minimal effect on humans.

CAPTCHA design seems to be both an art and a science. They will never be strong in the sense of a cryptosystem or a proof; by definition, they can be broken with nothing more than a few moments of casual human effort. In what may be an endless arms race, perhaps the best we can do is ensure the latest weapon *du jour* is fun for users. In this goal, at least, we hope Asirra succeeds.

7. REFERENCES

[1] Frozen Bear. hotcaptcha. http://www.hotcaptcha.com.

[2] Kumar Chellapilla, Kevin Larson, Patrice Simard, and Mary Czerwinski. Designing human friendly human interaction proofs (HIPs). In *Proceedings of ACM CHI 2005 Conference on Human Factors in Computing Systems*, volume 1 of *Email and security*, pages 711–720, 2005.

[3] Monica Chew and J.D. Tygar. Image recognition CAPTCHAs. In *Proceedings of the 7th International Information Security Conference (ISC 2004)*, pages 268–279. Springer, September 2004.

[4] Mark Everingham, Andrew Zisserman, Chris Williams, and Luc Van Gool. The PASCAL visual object classes challenge 2006 (VOC2006) results. Technical report, University of Oxford, 2006.

[5] Ralph Gross, Jianbo Shi, and Jeff Cohn. Quo vadis Face Recognition? Technical Report CMU-RI-TR-01-17, Carnegie Mellon University Robotics Institute, June 2001.

[6] Google Images. http://images.google.com.

[7] Eight Days Inc. Hot or not. http://www.hotornot.com.

[8] Greg Mori and Jitendra Malik. Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. In *Conference on Computer Vision and Pattern Recognition (CVPR '03)*, pages 134–144. IEEE Computer Society, 2003.

[9] Jared Saul. Petfinder. http://www.petfinder.com.

[10] Amazon Web Services. Ec2 scalable compute cloud. http://aws.amazon.com/ec2.

[11] Patrice Simard, David Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *International Conference on Document Analysis and Recognition*, pages 958–962. IEEE Computer Society, 2003.

[12] Digg.com user DoubtfulSalmon. http://tinyurl.com/2stwu3, April 2006.

[13] L. von Ahn, M. Blum, N.J. Hopper, and J. Langford. The CAPTCHA web page. http://www.captcha.net.

[14] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using hard AI problems for security. In Eli Biham, editor, *Advances in Cryptology - EUROCRYPT 2003, International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland, May 4-8, 2003, Proceedings*, volume 2656 of *Lecture Notes in Computer Science*, pages 294–311. Springer, 2003.

[15] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In Elizabeth Dykstra-Erickson and Manfred Tscheligi, editors, *Proceedings of the 2004 Conference on Human Factors in Computing Systems, CHI 2004, Vienna, Austria, April 24 - 29, 2004*, pages 319–326. ACM, 2004.

[16] Oli Warner. Kittenauth. http://www.thepcspy.com/kittenauth.

[17] WikiMedia Foundation. http://commons.wikimedia.org.

[18] Wen-Yi Zhao, Rama Chellappa, P. J. Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv*, 35(4):399–458, 2003.

PX0362

NuCaptcha Dynamically Alters Captchas To Promise Security

Leena Rao **August 11, 2011**

NuCaptcha, a startup that develops video captcha technology, is hoping to disrupt the captcha industry today with the launch of its dynamic captchas that promise security to website owners.

Captchas are security questions you find on Web sites that require you to decipher and type words or numbers and detects whether the user is a human or a spambot. Most Captchas you see are transcription, text-based Captchas.

Of course, captchas can often be confusing, and not user-friendly. NuCaptcha is trying to solve this pain point by offering site owners a more user-friendly alternative, called a video captcha. NuCaptcha has spent the last few years optimizing its security so that 3 un-distorted letters can be used on "most" sites. Sites that need more security can add up to 8 letters.

NuCaptcha's Behavioral Analysis System detects unusual behavior dynamically and can increase the complexity of the Captcha and slow down the video, making it difficult for both bots and human-farms to solve Captcha in high volumes.

A variety of video-based templates are available for site owners, based on their required site look-and-feel, and can be deployed quickly. Currently serving millions of Captchas per month, NuCaptcha offers a Free, Pro and Enterprise offering, and users can determine the size, choice of videos, security level, and even optional branding, advertising and revenue generation through NuCaptcha's Advertising Programs.

NuCaptcha faces competition from ReCaptcha, which Google acquired in 2009.

PX0363

PhotoDNA

Help stop the spread of child exploitation

In 2009, Microsoft partnered with Dartmouth College to develop PhotoDNA, a technology that aids in finding and removing known images of child exploitation. Today, PhotoDNA is used by organizations around the world and has assisted in the detection, disruption, and reporting of millions of child exploitation images.

Organizations, Businesses & Non-Profits

PhotoDNA Cloud Service is free for qualified customers. Learn more >

APPLY

Already a customer? Sign in. >

Law Enforcement

PhotoDNA is available for free to law enforcement and tool providers. Learn more >

How does PhotoDNA technology work?

PhotoDNA creates a unique digital signature (known as a "hash") of an image which is then compared against signatures (hashes) of other photos to find copies of the same image. When matched with a database containing hashes of previously identified illegal images, PhotoDNA is an incredible tool to help detect, disrupt and report the distribution of child exploitation material. PhotoDNA is not facial recognition software and cannot be used to identify a person or object in an image. A PhotoDNA hash is not reversible, and therefore cannot be used to recreate an image.

Who uses PhotoDNA technology?

Microsoft donated PhotoDNA to the National Center for Missing & Exploited Children (NCMEC). NCMEC is the United States clearinghouse and comprehensive reporting center for all issues related to the prevention of, and recovery from, child victimization, include abduction, abuse and exploitation. The CyberTipline provides the public and electronic service providers (ESPs) with the ability to report instances of online enticement of children for sexual acts and child exploitation material. For more information, or to register for the CyberTipline, visit NCMEC at **www.missingkids.com**.

Microsoft continues to provide this valuable technology for free to qualified organizations including technology companies, developers, and non-profit organizations for the purpose of combatting child exploitation. Microsoft has also provided PhotoDNA for free to law enforcement, primarily through forensic tool developers. PhotoDNA has been widely incorporated into innovative visual image and forensic tools used by law enforcement around the world.

In 2015, Microsoft made PhotoDNA available as a service on Azure. The PhotoDNA Cloud Service enables smaller companies and other organizations that want to give users the freedom to upload content while ensuring the integrity of their platforms.

"Our goal is to stop that victimization. Using PhotoDNA, we will be able to match those images, working with online service providers around the country, so we can stop the redistribution of the photos." – Ernie Allen, former President & CEO of the National Center for Missing & Exploited Children

"By working collaboratively with industry and sharing PhotoDNA technology, we're continuing the fight to help protect children." – Courtney Gregoire, Assistant General Counsel, Microsoft Digital Crimes Unit

PX0368

Welcome to Apache ZooKeeper™

Apache ZooKeeper is an effort to develop and maintain an open-source server which enables highly reliable distributed coordination.

What is ZooKeeper?

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.

Learn more about ZooKeeper on the ZooKeeper Wiki.

Getting Started

Start by installing ZooKeeper on a single machine or a very small cluster.

1. Learn about ZooKeeper by reading the documentation.
2. Download ZooKeeper from the release page.

Getting Involved

Apache ZooKeeper is an open source volunteer project under the Apache Software Foundation. We encourage you to learn about the project and contribute your expertise. Here are some starter links:

1. See our How to Contribute to ZooKeeper page.
2. Give us feedback: What can we do better?
3. Join the mailing list: Meet the community.

Copyright © 2010-2020 The Apache Software Foundation, Licensed under the Apache License, Version 2.0.
Apache ZooKeeper, ZooKeeper, Apache, the Apache feather logo, and the Apache ZooKeeper project logo are trademarks of The Apache Software Foundation.

PX0369

Apache Oozie Workflow Scheduler for Hadoop

Overview

Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions.

Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability.

Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).

Oozie is a scalable, reliable and extensible system.

Developers interested in getting more involved with Oozie may join the mailing lists, report bugs , retrieve code from the version control system, and make contributions .

Copyright © 2011-2021 The Apache Software Foundation. All Rights Reserved.

PX0370

The New York Times https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

The Silent Partner Cleaning Up Facebook for $500 Million a Year

The social network has constructed a vast infrastructure to keep toxic material off its platform. At the center of it is Accenture, the blue-chip consulting firm.

By Adam Satariano and Mike Isaac

Published Aug. 31, 2021 Updated Oct. 28, 2021

In 2019, Julie Sweet, the newly appointed chief executive of the global consulting firm Accenture, held a meeting with top managers. She had a question: Should Accenture get out of some of the work it was doing for a leading client, Facebook?

For years, tensions had mounted within Accenture over a certain task that it performed for the social network. In eight-hour shifts, thousands of its full-time employees and contractors were sorting through Facebook's most noxious posts, including images, videos and messages about suicides, beheadings and sexual acts, trying to prevent them from spreading online.

Some of those Accenture workers, who reviewed hundreds of Facebook posts in a shift, said they had started experiencing depression, anxiety and paranoia. In the United States, one worker had joined a class-action lawsuit to protest the working conditions. News coverage linked Accenture to the grisly work. So Ms. Sweet had ordered a review to discuss the growing ethical, legal and reputational risks.

At the meeting in Accenture's Washington office, she and Ellyn Shook, the head of human resources, voiced concerns about the psychological toll of the work for Facebook and the damage to the firm's reputation, attendees said. Some executives who oversaw the Facebook account argued that the problems were manageable. They said the social network was too lucrative a client to lose.

The meeting ended with no resolution.

Julie Sweet, the chief executive of Accenture, ordered a review of its business with Facebook. Greg Kahn for The New York Times

Facebook and Accenture have rarely talked about their arrangement or even acknowledged that they work with each other. But their secretive relationship lies at the heart of an effort by the world's largest social media company to distance itself from the most toxic part of its business.

For years, Facebook has been under scrutiny for the violent and hateful content that flows through its site. Mark Zuckerberg, the chief executive, has repeatedly pledged to clean up the platform. He has promoted the use of artificial intelligence to weed out toxic posts and touted efforts to hire thousands of workers to remove the messages that the A.I. doesn't.

But behind the scenes, Facebook has quietly paid others to take on much of the responsibility. Since 2012, the company has hired at least 10 consulting and staffing firms globally to sift through its posts, along with a wider web of subcontractors, according to interviews and public records.

No company has been more crucial to that endeavor than Accenture. The Fortune 500 firm, better known for providing high-end tech, accounting and consulting services to multinational companies and governments, has become Facebook's single biggest partner in moderating content, according to an examination by The New York Times.

Accenture has taken on the work — and given it a veneer of respectability — because Facebook has signed contracts with it for content moderation and other services worth at least $500 million a year, according to The Times's examination. Accenture employs more than a third of the 15,000 people whom Facebook has said it has hired to inspect its posts. And while the agreements provide only a small fraction of Accenture's annual revenue, they give it an important lifeline into Silicon Valley. Within Accenture, Facebook is known as a "diamond client."

Their contracts, which have not previously been reported, have redefined the traditional boundaries of an outsourcing relationship. Accenture has absorbed the worst facets of moderating content and made Facebook's content issues its own. As a cost of doing business, it has dealt with workers' mental health issues from

reviewing the posts. It has grappled with labor activism when those workers pushed for more pay and benefits. And it has silently borne public scrutiny when they have spoken out against the work.

Those issues have been compounded by Facebook's demanding hiring targets and performance goals and so many shifts in its content policies that Accenture struggled to keep up, 15 current and former employees said. And when faced with legal action from moderators about the work, Accenture stayed quiet as Facebook argued that it was not liable because the workers belonged to Accenture and others.

"You couldn't have Facebook as we know it today without Accenture," said Cori Crider, a co-founder of Foxglove, a law firm that represents content moderators. "Enablers like Accenture, for eye-watering fees, have let Facebook hold the core human problem of its business at arm's length."

The Times interviewed more than 40 current and former Accenture and Facebook employees, labor lawyers and others about the companies' relationship, which also includes accounting and advertising work. Most spoke anonymously because of nondisclosure agreements and fear of reprisal. The Times also reviewed Facebook and Accenture documents, legal records and regulatory filings.

Facebook and Accenture declined to make executives available for comment. Drew Pusateri, a Facebook spokesman, said the company was aware that content moderation "jobs can be difficult, which is why we work closely with our partners to constantly evaluate how to best support these teams."

Stacey Jones, an Accenture spokeswoman, said the work was a public service that was "essential to protecting our society by keeping the internet safe."

Neither company mentioned the other by name.

Pornographic Posts

Much of Facebook's work with Accenture traces back to a nudity problem.

In 2007, millions of users joined the social network every month — and many posted naked photos. A settlement that Facebook reached that year with Andrew M. Cuomo, who was New York's attorney general, required the company to take down pornographic posts flagged by users within 24 hours.

Facebook employees who policed content were soon overwhelmed by the volume of work, members of the team said. Sheryl Sandberg, the company's chief operating officer, and other executives pushed the team to find automated solutions for combing through the content, three of them said.

Mark Zuckerberg, Facebook's chief executive, has repeatedly pledged to clean up the platform and hailed efforts to hire thousands of moderators. Jessica Chou for The New York Times

Facebook also began looking at outsourcing, they said. Outsourcing was cheaper than hiring people and provided tax and regulatory benefits, along with the flexibility to grow or shrink quickly in regions where the company did not have offices or language expertise. Ms. Sandberg helped champion the outsourcing idea, they said, and midlevel managers worked out the details.

By 2011, Facebook was working with oDesk, a service that recruited freelancers to review content. But in 2012, after the news site Gawker reported that oDesk workers in Morocco and elsewhere were paid as little as $1 per hour for the work, Facebook began seeking another partner.

Facebook landed on Accenture. Formerly known as Andersen Consulting, the firm had rebranded as Accenture in 2001 after a break with the accounting firm Arthur Andersen. And it wanted to gain traction in Silicon Valley.

In 2010, Accenture scored an accounting contract with Facebook. By 2012, that had expanded to include a deal for moderating content, particularly outside the United States.

That year, Facebook sent employees to Manila and Warsaw to train Accenture workers to sort through posts, two former Facebook employees involved with the trip said. Accenture's workers were taught to use a Facebook software system and the platform's guidelines for leaving content up, taking it down or escalating it for review.

'Honey Badger'

What started as a few dozen Accenture moderators grew rapidly.

By 2015, Accenture's office in the San Francisco Bay Area had set up a team, code-named Honey Badger, just for Facebook's needs, former employees said. Accenture went from providing about 300 workers in 2015 to about 3,000 in 2016. They are a mix of full-time employees and contractors, depending on the location and task.

The firm soon parlayed its work with Facebook into moderation contracts with YouTube, Twitter, Pinterest and others, executives said. (The digital content moderation industry is projected to reach $8.8 billion next year, according to Everest Group, roughly double the 2020 total.) Facebook also gave Accenture contracts in areas like checking for fake or duplicate user accounts and monitoring celebrity and brand accounts to ensure they were not flooded with abuse.

After federal authorities discovered in 2016 that Russian operatives had used Facebook to spread divisive posts to American voters for the presidential election, the company ramped up the number of moderators. It said it would hire more than 3,000 people — on top of the 4,500 it already had — to police the platform.

“If we’re going to build a safe community, we need to respond quickly,” Mr. Zuckerberg said in a 2017 post.

The next year, Facebook hired Arun Chandra, a former Hewlett Packard Enterprise executive, as vice president of scaled operations to help oversee the relationship with Accenture and others. His division is overseen by Ms. Sandberg.

Facebook also spread the content work to other firms, such as Cognizant and TaskUs. Facebook now provides a third of TaskUs’s business, or $150 million a year, according to regulatory filings.

The work was challenging. While more than 90 percent of objectionable material that comes across Facebook and Instagram is removed by A.I., outsourced workers must decide whether to leave up the posts that the A.I. doesn’t catch.

They receive a performance score that is based on correctly reviewing posts against Facebook’s policies. If they make mistakes more than 5 percent of the time, they can be fired, Accenture employees said.

But Facebook’s rules about what was acceptable changed constantly, causing confusion. When people used a gas-station emoji as slang for selling marijuana, workers deleted the posts for violating the company’s content policy on drugs. Facebook then told moderators not to remove the posts, before later reversing course.

Facebook also tweaked its moderation technology, adding new keyboard shortcuts to speed up the review process. But the updates were sometimes released with little warning, increasing errors.

As of May, Accenture billed Facebook for roughly 1,900 full-time moderators in Manila; 1,300 in Mumbai, India; 850 in Lisbon; 780 in Kuala Lumpur, Malaysia; 300 in Warsaw; 300 in Mountain View, Calif.; 225 in Dublin; and 135 in Austin, Texas, according to staffing records reviewed by The Times.

At the end of each month, Accenture sent invoices to Facebook detailing the hours worked by its moderators and the volume of content reviewed. Each U.S. moderator generated $50 or more per hour for Accenture, two people with knowledge of the billing said. In contrast, moderators in some U.S. cities received starting pay of $18 an hour.

Psychological Costs

Within Accenture, workers began questioning the effects of viewing so many hateful posts.

Accenture hired mental health counselors to handle the fallout. Izabela Dziugiel, a counselor who worked in Accenture's Warsaw office, said she told managers in 2018 that they were hiring people ill prepared to sort through the content. Her office handled posts from the Middle East, including gruesome images and videos of the Syrian war.

"They would just hire anybody," Izabela Dziugiel, a former mental health counselor for Accenture, said of the firm. Zuza Krajewska for The New York Times

"They would just hire anybody," said Ms. Dziugiel, who previously treated soldiers with post-traumatic stress disorder. She left the firm in 2019.

In Dublin, one Accenture moderator who sifted through Facebook content left a suicide note on his desk in 2018, said a mental health counselor who was involved in the episode. The worker was found safe.

Joshua Sklar, a moderator in Austin, who quit in April, said he had reviewed 500 to 700 posts a shift, including images of dead bodies after car crashes and videos of animals being tortured.

“One video that I watched was a guy who was filming himself raping a little girl,” said Mr. Sklar, who described his experience in an internal post that later became public. “It was just awful.”

If workers went around Accenture’s chain of command and directly communicated with Facebook about content issues, they risked being reprimanded, he added. That made Facebook slower to learn about and react to problems, he said.

Facebook said anyone filtering content could escalate concerns.

Another former moderator in Austin, Spencer Darr, said in a legal hearing in June that the job had required him to make unimaginable decisions, such as whether to delete a video of a dog being skinned alive or simply mark it as disturbing. “Content moderators’ job is an impossible one,” he said.

Joshua Sklar, a moderator who quit in April, said he had reviewed 500 to 700 posts a shift, including images of dead bodies after car crashes. Lauren Withrow for The New York Times

In 2018, Accenture introduced WeCare — policies that mental health counselors said limited their ability to treat workers. Their titles were changed to "wellness coaches" and they were instructed not to give psychological assessments or diagnoses, but to provide "short-term support" like taking walks or listening to calming music. The goal, according to a 2018 Accenture guidebook, was to teach moderators "how to respond to difficult situations and content."

Accenture's Ms. Jones said the company was "committed to helping our people who do this important work succeed both professionally and personally." Workers can see outside psychologists.

By 2019, scrutiny of the industry was growing. That year, Cognizant said it was exiting content moderation after the tech site The Verge described the low pay and mental health effects of workers at an Arizona office. Cognizant said the decision would cost it at least $240 million in revenue and lead to 6,000 job cuts.

Internal Debate

More than one Accenture chief executive debated doing business with Facebook.

In 2017, Pierre Nanterme, Accenture's chief at the time, questioned the ethics of the work and whether it fit the firm's long-term strategy of providing services with high profit margins and technical expertise, three executives involved in the discussions said.

No actions were taken. Mr. Nanterme died from cancer in January 2019.

Five months later, Ms. Sweet, a longtime Accenture lawyer and executive, was named chief executive. She soon ordered the review of the moderation business, three former colleagues said.

Executives prepared reports and debated how the work compared with jobs like an ambulance driver. Consultants were sent to observe moderators and their managers.

The office in Austin, which had opened in 2017, was selected for an audit as part of Ms. Sweet's review. The city was also home to a Facebook office and had large populations of Spanish and Arabic speakers to read non-English posts. At its peak, Accenture's Austin office had about 300 moderators parsing through Facebook posts.

But some workers there became unhappy about the pay and viewing so much toxic content. Organizing through text messages and internal message boards, they called for better wages and benefits. Some shared their stories with the media.

The office building in Austin, Texas, where Accenture moderators who sorted content for Facebook worked. Lauren Withrow for The New York Times

Last year, a worker in Austin was one of two from Accenture who joined a class-action suit against Facebook filed by U.S. moderators. Facebook argued that it was not liable because the workers were employed by firms like Accenture, according to court records. The social network reached a $52 million settlement with the workers in May 2020.

For Ms. Sweet, the debate over the Facebook contracts stretched out over several meetings, former executives said. She subsequently made several changes.

In December 2019, Accenture created a two-page legal disclosure to inform moderators about the risks of the job. The work had “the potential to negatively impact your emotional or mental health,” the document said.

Last October, Accenture went further. It listed content moderation for the first time as a risk factor in its annual report, saying it could leave the firm vulnerable to media scrutiny and legal trouble. Accenture also restricted new moderation clients, two people with knowledge of the policy shift said. Any new contracts required approval from senior management.

But Ms. Sweet also left some things untouched, they said.

Among them: the contracts with Facebook. Ultimately, the people said, the client was too valuable to walk away from.

Adam Satariano is a technology reporter based in London. More about Adam Satariano

Mike Isaac is a technology correspondent and the author of Super Pumped: The Battle for Uber, a NYT best-selling book on the dramatic rise and fall of the ride-hailing company. He regularly covers Facebook and Silicon Valley, and is based in The Times's San Francisco bureau. More about Mike Isaac

A version of this article appears in print on , Section A, Page 1 of the New York edition with the headline: Profit and Pain in Doing Facebook's Dirty Work

PX0376

Meta

Resources

Models and libraries

Our open source libraries and models for those taking our AI learnings further through software and app development

Meta

Models Libraries

Models

Computer vision

Detectron 2

Our next-generation platform for object detection and segmentation

Overview

DensePose

Designed to map all human pixels of an RGB image to a 3D surface-based representation of the human body

Overview

Language

Seamless

SeamlessM4T is a foundational speech/text translation and transcription model that overcomes the limitations of previous systems with state-of-the-art results.

Overview

Llama

Shaping the next wave of innovation through access of Llama's open platform featuring AI models, tools, and resources.

Overview

Fairseq

Meta

Overview

VizSeq

A research toolkit for natural language generation (translation, captioning, summarization, and more)

Overview

Multimodal

AudioCraft

A single codebase for developing audio generative models.

Overview

Hateful Memes Dataset

An open source dataset designed for the development of new systems that identify multimodal hate speech.

Overview

Reasoning

ELF

A platform for game research that allows developers to train and test their algorithms in various game environments

Overview

ELF OpenGo

An AI bot from Fundamental AI Research (FAIR) that has defeated world champion professional Go players

Overview

Libraries

Meta

Language

FastText

A lightweight library designed to help build scalable solutions for text representation and classification

Overview

Translate

An open source project based on AI at Meta's machine translation systems

Overview

MUSE

A Python library that enables faster development and evaluation of cross-lingual word embeddings and NLP

Overview

CoVoST

A large-scale multilingual speech-to-text translation corpus

Overview

FAISS

Allows developers to quickly search for embeddings of multimedia documents that are similar to each other

Overview

Reasoning

ReAgent

A platform for Large-Scale Reasoning systems (Reinforcement Learning, Contextual Bandits)

Meta

PHYRE

Benchmark for PHYsical REasoning that contains a set of simple classical mechanics puzzles in a 2D environment

Overview

TorchCraft

Enables AI research on real-time strategy (RTS) games such as StarCraft: Brood War

Overview

Other resources

Explore more of our open science resources

From demos and systems cards to frameworks and tools, we offer a host of resources to help you accelerate AI your way.

1 / 5

Frameworks and tools

Sharing our ML frameworks and tools with the community to collaborate and

Datasets

Large-scale datasets and benchmarks for training, evaluating, and testing models to

Demos

Our demos for anyone wanting experience our latest research

Meta

Learn more Learn more Learn more

Search AI content

Who We Are

Latest Work

Our Actions

Newsletter

Privacy Policy
Terms
Cookies

Meta © 2024

PX0377

Meta

ML Applications | Integrity

How Facebook uses super-efficient AI models to detect hate speech

November 19, 2020

Share on Facebook

Share on Twitter

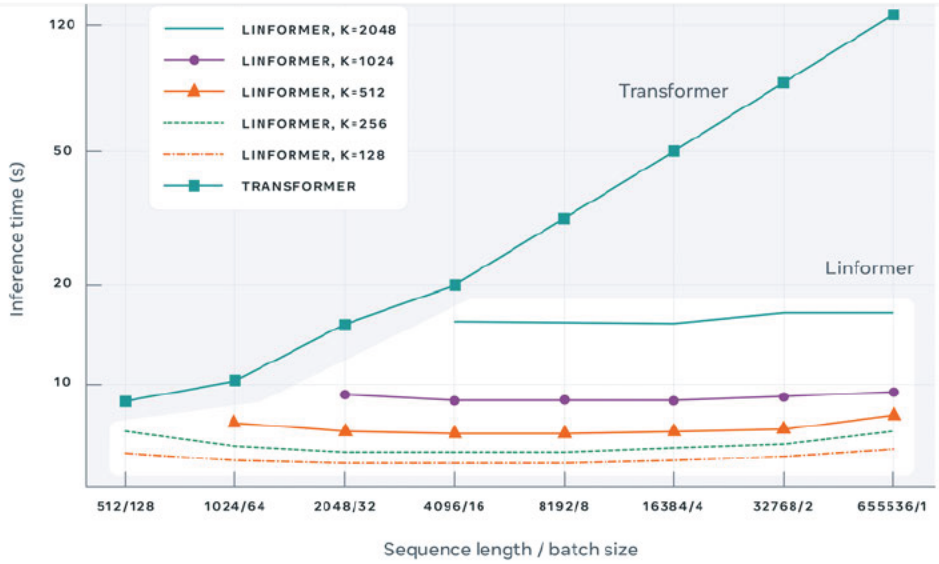
Meta

Building AI that can analyze complicated text isn't enough to protect people from harmful content. We need systems that spot a slang-filled or intentionally misspelled piece of hate speech — and do it in a fraction of a second and at billion-person scale.

This has been a challenge when deploying systems to detect hate speech because the most powerful, cutting-edge language-understanding systems today use large-scale Transformer models with hundreds of millions or billions of parameters. New models, including Facebook AI's RoBERTa and XLM-R, have repeatedly advanced the state of the art, but these gains have come from creating ever-larger models that require massive amounts of computation.

To unlock the capabilities of these powerful AI models, Facebook AI recently developed a new Transformer architecture called Linformer. It makes it possible to use them efficiently at scale. Linformer is the first theoretically proven linear-time Transformer architecture. With standard Transformers, the amount of required processing power increases at a geometric rate as the input length increases. With Linformer, however, the number of computations increases only at a linear rate. This makes it possible to use larger pieces of text to train models, and thereby achieve better performance.

Meta



We are now using Linformer to analyze billions of pieces of content on Facebook and Instagram in different regions around the world.

	TRANSFORMER (2017)	SPARSE TRANSFORMER (2019)	REFORMER (2020)	LINFORMER (2020)
Complexity / layer	O(N2)	O(N1.5)	O(NlogN)	O(N)
Sequential operation	O(1)	O(1)	O(logN)	O(1)

This chart compares the level of complexity of different Transformer architectures.

Meta

little of the hate speech on our platforms that we removed was done so before anyone reported it. As detailed in Facebook's quarterly Community Standards Enforcement Report released today, AI proactively detected 94.7 percent of the hate speech we removed.

Earlier this year, we published our research on Linformer and released our code so other researchers and engineers could improve their models. Since our Facebook AI Research (FAIR) lab was founded in 2013, we've committed to an open science-based approach. Our research model revolves around publishing code and methodologies, collaborating with other researchers across industry and academia, and creating open benchmarks and challenges. We're now sharing details here on how Linformer works and how we are using it to keep people safe on our platforms.

These are difficult problems, and our systems are still far from perfect. And even if we had perfect AI tools, there would still be difficult questions about what policies will serve people best. But progress in AI has made our platforms better and safer, and we are working hard to advance our technology further.

A simpler way to build cutting-edge AI models

Meta

reinforcement learning. Transformers rely on a simple-yet-powerful mechanism called self-attention, which enables AI models to selectively focus on certain parts of their input and thus better understand the content.

They are notoriously resource-intensive, however, because their self-attention mechanism requires more and more memory and computations as the length of the input sequence grows. When you increase the input size from, say, to 4,000 to 8,000, the number of computations doesn't double. It goes from about 16,000,000 to 64,000,000.

Linformer overcomes this challenge by approximating the information in the attention matrix — without degrading the model's performance. As shown in the chart above, Linformer models can make predictions efficiently even as the size of the input grows.

This graphic shows how Linformer produces lower-rank matrices.

To build Linformer, we first demonstrated that low-rank approximation is possible for self-attention. This indicated it was possible to greatly simplify the standard Transformer architecture without degrading performance.

Meta

This flowchart shows how Linformer can be used to create models to detect hate speech

Time saved

Length (n)	PROJECTED DIMENSIONS (K) 128	256	512	1024	2048
512	1.5x	1.3x	-	-	-
1024	1.7x	1.6x	1.3x	-	-
2048	2.6x	2.4x	2.1x	1.3x	-
4096	3.4x	3.2x	2.8x	2.2x	1.3x
8192	5.5x	5.0x	4.4x	3.5x	2.1x
16384	8.6x	7.8x	7.0x	5.6x	3.3x
32768	13x	12x	11x	8.8x	5.0x
65536	20x	18x	16x	14x	7.9x

Memory saved

Length (n)	PROJECTED DIMENSIONS (K) 128	256	512	1024	2048
512	1.7x	1.5x	-	-	-
1024	3.0x	2.9x	1.8x	-	-
2048	6.1x	5.6x	3.6x	2.0x	-
4096	14x	13x	8.3x	4.3x	2.3x
8192	28x	26x	17x	8.5x	4.5x
16384	56x	48x	32x	16x	8x
32768	56x	48x	36x	18x	16x
65536	60x	52x	40x	20x	18x

As the sequence length increases, Linformer's efficiency gains grow.

In a typical Transformer, every single token at each layer must look at (or attend to) every other single token from the previous layer. This generates quadratic

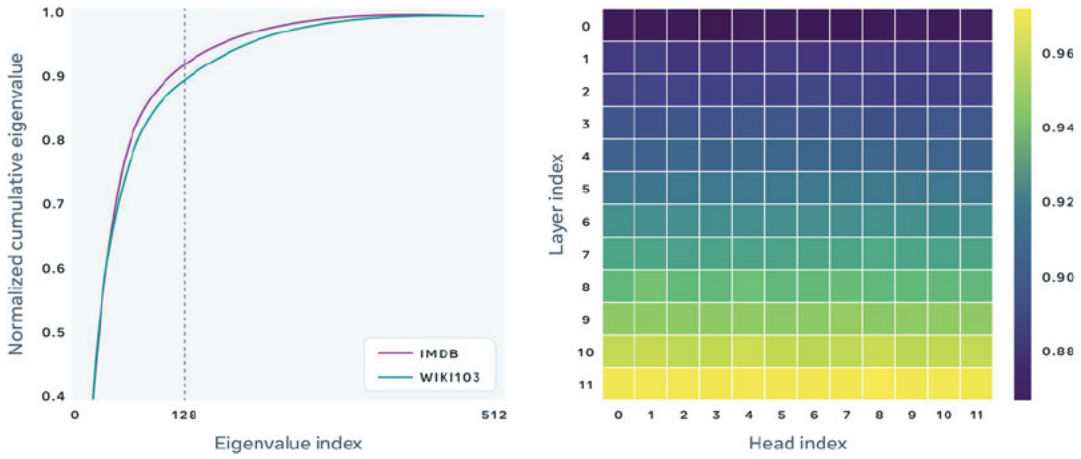
Meta

N	MODEL	SST-2	IMDB	QNLI	QQP	AVERAGE
512	Liu et al. (2019), RoBERTa-base	93.1	94.1	90.9	**90.9**	92.25
	Linformer, 128	92.4	94.0	90.4	90.2	91.75
	Linformer, 128, shared kv	**93.4**	93.4	90.3	90.3	91.85
	Linformer, 128, shared kv, layer	93.2	93.8	90.1	90.2	91.83
	Linformer, 256	93.2	94.0	90.6	90.5	92.08
	Linformer, 256, shared kv	93.3	93.6	90.6	90.6	92.03
	Linformer, 256, shared kv, layer	93.1	94.1	**91.2**	90.8	**92.30**
512	Devlin et al. (2019), BERT-base	92.7	93.5	91.8	89.6	91.90
	Sanh et al. (2019), Distilled BERT	91.3	92.8	89.2	88.5	90.45
1024	Linformer, 256	93.0	93.8	90.4	90.4	91.90
	Linformer, 256, shared kv	93.0	93.6	90.3	90.4	91.83
	Linformer, 256, shared kv, layer	93.2	**94.2**	90.8	90.5	92.18

This chart compares the performance of Linformer models with other Transformer architectures.

We tested two large datasets, WIKI103 and IMDB, with a RoBERTa Transformer model in order to calculate the eigenvalues, a standard measurement of the approximate rank of the matrix.

We were able to demonstrate that the information from N tokens of the previous layer can be compressed into a smaller, fixed-size set of K distinct units. With this compression, the system needs to iterate only across this smaller set of K units for each token. In other words, self-attention is what is known as low-rank and can be expressed using a smaller matrix of numbers.



Spectrum analysis of the self-attention matrix of Transformer model.

What's next

These efficiency gains matter because we want to deal with hate speech before it has a chance to spread. Milliseconds count when determining whether a post violates our policies. Linformer is helping us do this today, but it could also lead to new AI-powered integrity systems that simply aren't currently possible. Could we one day deploy a state-of-the-art model that learns from text, images, and speech and effectively detects not just hate speech but human trafficking, bullying, and other forms of harmful content?

There is much work to do before we reach this goal, but Linformer brings us one step closer. What's more, by making large-scale Transformer models more efficient, Linformer also enables smaller research labs and engineering teams to train and test state-of-the-art AI even if they don't have access to massive computing resources.

We're committed to keeping people safe on our platforms. We believe that Linformer and other AI advances will enable us to continue making progress, and

Meta

Privacy Policy
Terms
Cookies

Meta © 2024

PX0378

VentureBeat

Pinterest details the AI that powers its content moderation

Kyle Wiggers

@Kyle_L_Wiggers

March 5, 2021 8:00 AM

Image Credit: Reuters

Join us in Atlanta on April 10th and explore the landscape of security workforce. We will explore the vision, benefits, and use cases of AI for security teams. Request an invite here.

Pinterest this morning peeled back the curtains on the AI and machine learning technologies it's using to combat harmful content on its platform. Leveraging algorithms to automatically detect adult content, hateful activities, medical misinformation, drugs, graphic violence, and more before it's reported, the company says that policy-violating reports per impression have declined by 52% since fall 2019, when the technologies were first introduced. And reports for self-harm content have decreased by 80% since April 2019.

One of the challenges in building multi-category machine learning models for content safety is the scarcity of labeled data, forcing engineers to use simpler models that can't be extended to multi-model inputs. Pinterest solves this problem with a system trained on millions of human-reviewed Pins, consisting of both user reports and proactive model-based sampling from its Trust and Safety operations team, which assigns categories and takes action on violating content. The company also employs a Pin model trained using a mathematical, model-friendly representation of Pins based on their

keywords and images, aggregated with another model to generate scores that indicate which Pinterest boards might be in violation.

category [model] using multi-modal inputs — embeddings and text — for content safety is a valuable insight for decision makers ... We use a combination of offline and online models to get both performance and speed, providing a system design that's a nice learning for others and generally applicable."

ADVERTISEMENT

VB EVENT

The AI Impact Tour – Atlanta

Continuing our tour, we're headed to Atlanta for the AI Impact Tour stop on April 10th. This exclusive, invite-only event, in partnership with Microsoft, will feature discussions on how generative AI is transforming the security workforce. Space is limited, so request an invite today.

Request an invite

In production, Pinterest employs a family of models to proactively detect policy-violating Pins. When enforcing policies across Pins, the platform groups together Pins with similar images and identifies them by a unique hash called "image-signature." Models generate scores for each image-signature, and based on these scores, the same content moderation decision is applied to all Pins with the same image-signature.

ADVERTISEMENT

For example, one of Pinterest's models identifies Pins that it believes violates the platform's policy on health misinformation. Trained using labels from Pinterest, the

model internally finds keywords or text associated with misinformation and blocks pins with that language while at the same time identifying visual representations associated with medical misinformation. It accounts for factors like image and URL and blocks any images online across Pinterest search, the home feed, and related pins, according to Singh.

Since users usually save thematically related Pins together as a collection on boards around topics like recipes, Pinterest deployed a machine learning model to produce scores for boards and enforce board-level moderation. A Pin model trained using only embeddings — i.e., representations — generates content safety scores for each Pinterest board. An embedding for the boards is constructed by aggregating the embeddings of the most recent Pins saved to them. When fed into the Pin model, these embeddings produce a content safety score for each board, allowing Pinterest to identify policy-violating boards without training a model for boards.

"These technologies, along with an algorithm that rewards positive content, and policy and product updates such as blocking anti-vaccination content, prohibiting culturally insensitive ads, prohibiting political ads, and launching compassionate search for mental wellness, are the foundation for making Pinterest an inspiring place online," Singh said. "Our work has demonstrated the impact graph convolutional methods can have in a production recommender systems, as well as other graph representation learning problems at large scale, including knowledge graph reasoning and graph clustering."

VB Daily

Stay in the know! Get the latest news in your inbox daily

Email Subscribe

By subscribing, you agree to VentureBeat's Terms of Service.

Press Releases Contact Us Advertise Share a News Tip

Contribute to DataDecisionMakers

Privacy Policy Terms of Service Do Not Sell My Personal Information

© 2024 VentureBeat. All rights reserved.

PX0380

Perspective

How it Works

Using Machine Learning to Reduce Toxicity Online

Perspective uses machine learning models to identify abusive comments. The models score a phrase based on the perceived impact the text may have in a conversation. Developers and publishers can use this score to give feedback to commenters, help moderators more easily review comments, or help readers filter out "toxic" language.

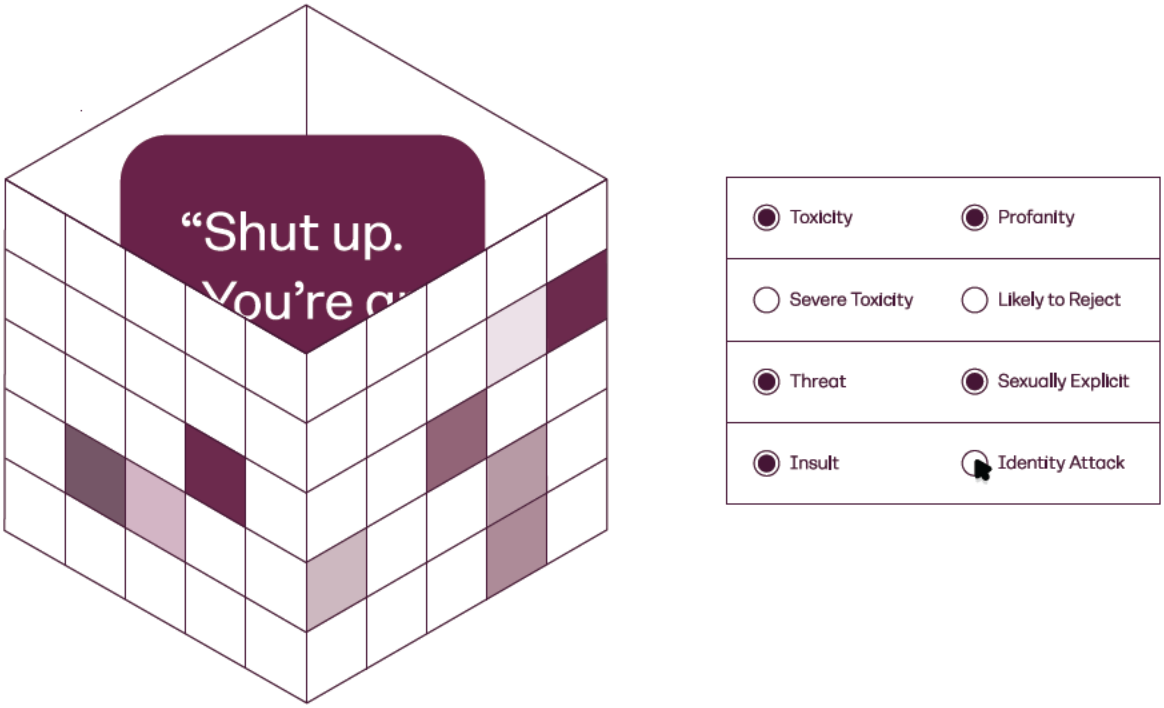
Perspective models provide scores for several different attributes. In addition to the flagship Toxicity attribute, here are some of the other attributes Perspective can provide scores for:

- Severe Toxicity
- Insult
- Profanity
- Identity attack
- Threat
- Sexually explicit

To learn more about our ongoing research and experimental models, visit our Developers site.

LEARN MORE

Perspective

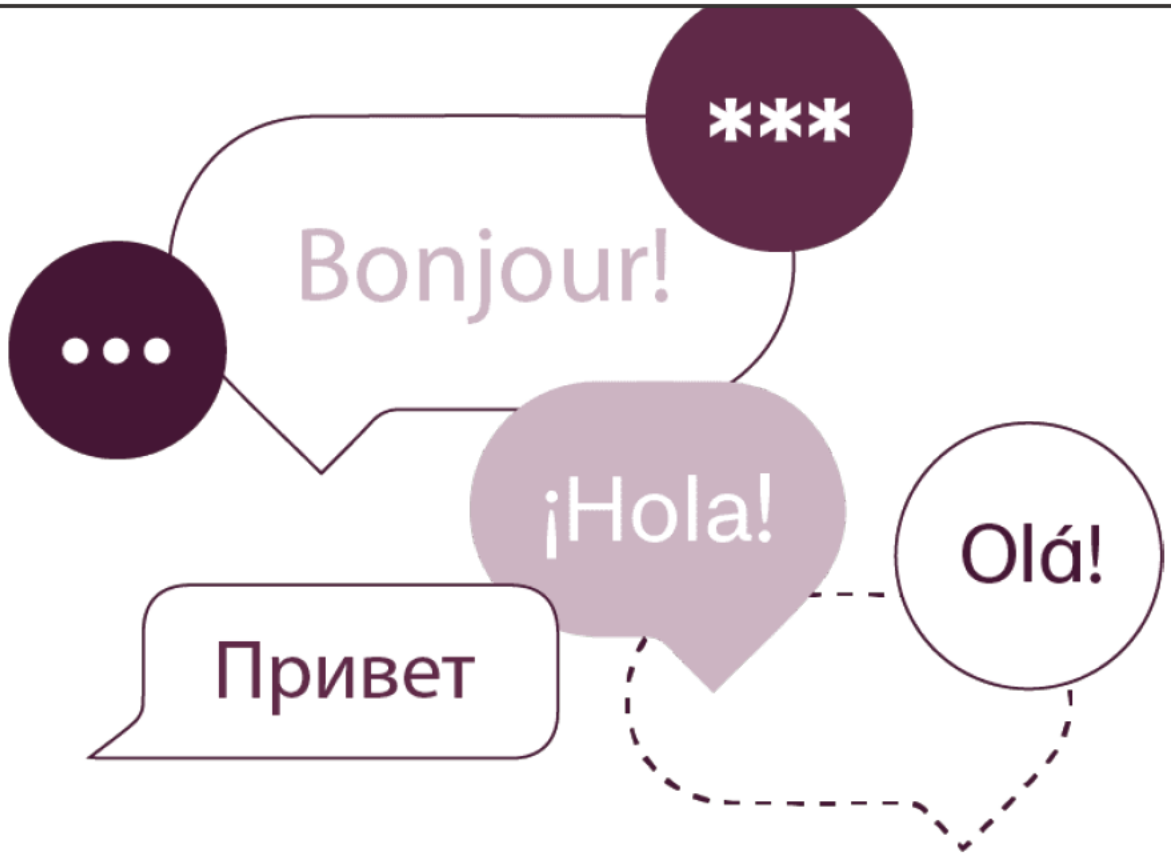


Language availability

Perspective API is free and available to use in Arabic, Chinese, Czech, Dutch, English, French, German, Hindi, Hinglish, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, and Swedish. The team is constantly developing models to support new languages.

To learn more about international publishers and platforms using Perspective API, check out our case studies. To learn more about languages in development, visit the Developers site.

Perspective



Perspective

Applying the model

Perspective in Action

One of Perspective API's most common uses is content moderation. Developers at platforms like DISQUS and publications like The Wall Street Journal augment their existing content moderation systems with Perspective, creating custom combinations of attributes to suit the needs of their platforms.

Perspective API is meant to make content moderation less of a burden on individuals and organizations, but is not meant to completely replace the work of human decision-makers. Perspective API helps developers and moderators make better decisions at scale, allowing healthy dialogue to flourish.

Perspective

How to customize Perspective to your needs

1. Choose the combination of attributes you'd like to use
2. Choose which language you'd like to support
3. Determine what actions should be triggered based on the output

Looking to learn more? Visit our Developers site for more technical information.

GO TO DEVELOPERS SITE

Perspective

Perspective API is the product of a collaborative research effort by Jigsaw and Google's Counter Abuse Technology team. We open source experiments, tools, and research data that explore ways to combat online toxicity and harassment.

English (United States)

Perspective Developers

Medium

Jigsaw

Contact Us

Privacy - Site Terms

API Terms

2021 © All rights reserved

PX0388

Meta

Research

RoBERTa: An optimized method for pretraining self-supervised NLP systems

7/29/2019

Share on Facebook

Share on Twitter

Meta

A robustly optimized method for pretraining natural language processing (NLP) systems that improves on Bidirectional Encoder Representations from Transformers, or BERT, the self-supervised method released by Google in 2018. BERT is a revolutionary technique that achieved state-of-the-art results on a range of NLP tasks while relying on unannotated text drawn from the web, as opposed to a language corpus that's been labeled specifically for a given task. The technique has since become popular both as an NLP research baseline and as a final task architecture. BERT also highlights the collaborative nature of AI research — thanks to Google's open release, we were able to conduct a replication study of BERT, revealing opportunities to improve its performance. Our optimized method, RoBERTa, produces state-of-the-art results on the widely used NLP benchmark, General Language Understanding Evaluation (GLUE).

In addition to a paper detailing those results, we're releasing the models and code that we used to demonstrate our approach's effectiveness.

How it works:

RoBERTa builds on BERT's language masking strategy, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples. RoBERTa, which was implemented in PyTorch, modifies key hyperparameters in BERT, including removing BERT's next-sentence pretraining objective, and training with much larger mini-batches and learning rates. This allows RoBERTa to improve on the masked language modeling objective compared with BERT and leads to better downstream task performance. We also explore training RoBERTa on an order of magnitude more data than BERT, for a longer amount of time. We used existing unannotated NLP datasets as well as CC-News, a novel set drawn from public news articles.

After implementing these design changes, our model delivered state-of-the-art performance on the MNLI, QNLI, RTE, STS-B, and RACE tasks and a sizable performance improvement on the GLUE benchmark. With a score of 88.5, RoBERTa reached the top position on the GLUE leaderboard, matching the performance of the previous leader, XLNet-Large. These results highlight the importance of previously unexplored design choices in BERT training and help disentangle the relative contributions of data size, training time, and pretraining objectives.

Meta

Our results show that tuning the BERT training procedure can significantly improve its performance on a variety of NLP tasks while also indicating that this overall approach remains competitive with alternative approaches. More broadly, this research further demonstrates the potential for self-supervised training techniques to match or exceed the performance of more traditional, supervised approaches. RoBERTa is part of Facebook's ongoing commitment to advancing the state-of-the-art in self-supervised systems that can be developed with less reliance on time- and resource-intensive data labeling. We look forward to seeing what the wider community does with the model and code for RoBERTa.

Read the full paper:

RoBERTa: A robustly optimized BERT pretraining approach

Research Areas

Research

Natural Language Processing

Related Posts

Meta

Research

Q&A: Facebook AI Residency Program

March 15, 2019

Meta

Hardware | Research

Yann LeCun on the future of deep learning hardware

February 18, 2019

Search AI content

Meta

Latest Work

Our Actions

Newsletter

Privacy Policy
Terms
Cookies

Meta © 2024

PX0389

Meta

NLP | Research

XLM-R: State-of-the-art cross-lingual understanding through self-supervision

11/7/2019

Share on Facebook

Share on Twitter

What the research is:

A new model, called XLM-R, that uses self-supervised training techniques to achieve state-of-the-art performance in cross-lingual understanding, a task in which a model is trained in one language and then used with other languages without additional training data. Our model improves upon previous multilingual approaches by incorporating more training data and languages — including so-called low-resource languages, which lack extensive labeled and unlabeled datasets.

XLM-R has achieved the best results to date on four cross-lingual understanding benchmarks, with increases of 4.7 percent average accuracy on the XNLI cross-lingual natural language inference dataset, 8.4 percent average F1 score on the recently introduced MLQA question answering dataset, and 2.1 percent F1 score on

Meta

rely on pretrained models.

In addition to sharing our results, we're releasing the code and models that we used for this research. Those resources can be found on our fairseq, Pytext and XLM repositories on GitHub.

How it works:

While earlier work in this area has demonstrated the effectiveness of multilingual masked language models on cross-lingual understanding, models such as XLM and multilingual BERT were limited in their ability to learn useful representations for low-resource languages. XLM-R improves on previous approaches in several ways:

- Building on the cross-lingual approach that we used with XLM and RoBERTa, we increased the number of languages and training examples for our new model, training self-supervised cross-lingual representations from more than two terabytes of publicly available CommonCrawl data that had been cleaned and filtered. This included generating new unlabeled corpora for low-resource languages, scaling the amount of training data available for those languages by two orders of magnitude.
- During fine-tuning, we leveraged the ability of multilingual models to use labeled data in multiple languages in order to improve downstream task performance. This enabled our model to achieve state-of-the-art results on cross-lingual benchmarks while exceeding the per-language performance of monolingual BERT models.
- We tuned our model's parameters to offset the fact that using cross-lingual transfer to scale models to more languages also limits the model's capacity to understand each of those languages. Our parameter changes included upsampling low-resource languages during training and vocabulary construction, generating a larger shared vocabulary, and increasing the overall model capacity up to 550 million parameters.

We found that XLM-R performed particularly well for low-resource languages, improving XNLI performance on Swahili and Urdu by 2.3 percent and 5 percent

Meta

Why it matters:

With people on Facebook posting content in more than 160 languages, XLM-R represents an important step toward our vision of providing the best possible experience on our platforms for everyone, regardless of what language they speak. Potential applications include serving highly accurate models for identifying hate speech and other policy-violating content across a wide range of languages. As this work helps us transition toward a one-model-for-many-languages approach — as opposed to one model per language — it will also make it easier to continue launching high-performing products in multiple languages at once. And by open-sourcing our models and code, (available through our GitHub repositories for fairseq, Pytext and XLM) we hope to improve the performance of multilingual models created by the research community, particularly systems that use self-supervised training methods to better understand low-resource languages.

Read the full paper:

Unsupervised Cross-lingual Representation Learning at Scale

Research Areas

Research

Natural Language Processing

Related Posts

Meta

Research

Q&A: Facebook AI Residency Program

March 15, 2019

Meta

Hardware | Research

Yann LeCun on the future of deep learning hardware

February 18, 2019

Search AI content

Meta

Latest Work

Our Actions

Newsletter

Privacy Policy
Terms
Cookies

Meta © 2024

PX0391

Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing

FRIDAY, NOVEMBER 02, 2018

Posted by Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language

One of the biggest challenges in natural language processing (NLP) is the shortage of training data. Because NLP is a diversified field with many distinct tasks, most task-specific datasets contain only a few thousand or a few hundred thousand human-labeled training examples. However, modern deep learning-based NLP models see benefits from much larger amounts of data, improving when trained on millions, or *billions*, of annotated training examples. To help close this gap in data, researchers have developed a variety of techniques for training general purpose language representation models using the enormous amount of unannotated text on the web (known as *pre-training*). The pre-trained model can then be *fine-tuned* on small-data NLP tasks like question answering and sentiment analysis, resulting in substantial accuracy improvements compared to training on these datasets from scratch.

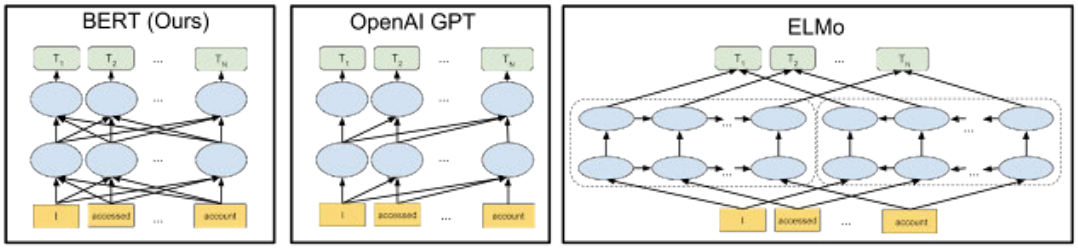
This week, we open sourced a new technique for NLP pre-training called **B**idirectional **E**ncoder **R**epresentations from Transformers, or **BERT**. With this release, anyone in the world can train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU. The release includes source code built on top of TensorFlow and a number of pre-trained language representation models. In our associated paper, we demonstrate state-of-the-art results on 11 NLP tasks, including the very competitive Stanford Question Answering Dataset (SQuAD v1.1).

What Makes BERT Different?

BERT builds upon recent work in pre-training contextual representations — including Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, and ULMFit. However, unlike these previous models, BERT is the first *deeply bidirectional*, *unsupervised* language representation, pre-trained using only a plain text corpus (in this case, Wikipedia).

models such as **word2vec** or **GloVe** generate a single **word embedding** representation for each word in the vocabulary. For example, the word "*bank*" would have the same context-free representation in "*bank account*" and "*bank of the river.*" Contextual models instead generate a representation of each word that is based on the other words in the sentence. For example, in the sentence "*I accessed the bank account*," a unidirectional contextual model would represent "*bank*" based on "*I accessed the*" but not "*account*." However, BERT represents "*bank*" using both its previous and next context — "*I accessed the ... account*" — starting from the very bottom of a deep neural network, making it deeply bidirectional.

A visualization of BERT's neural network architecture compared to previous state-of-the-art contextual pre-training methods is shown below. The arrows indicate the information flow from one layer to the next. The green boxes at the top indicate the final contextualized representation of each input word:



BERT is deeply bidirectional, OpenAI GPT is unidirectional, and ELMo is shallowly bidirectional.

The Strength of Bidirectionality

If bidirectionality is so powerful, why hasn't it been done before? To understand why, consider that unidirectional models are efficiently trained by predicting each word conditioned on the previous words in the sentence. However, it is not possible to train bidirectional models by simply conditioning each word on its previous *and* next words, since this would allow the word that's being predicted to indirectly "see itself" in a multi-layer model.

To solve this problem, we use the straightforward technique of masking out some of the words in the input and then condition each word bidirectionally to predict the masked words. For example:

Input: The man went to the $[MASK]_1$. He bought a $[MASK]_2$ of milk .
Labels: $[MASK]_1$ = store; $[MASK]_2$ = gallon

BERT also learns to model relationships between sentences by pre-training on a very simple task that can be generated from any text corpus: Given two sentences `A` and `B`, is `B` the actual next sentence that comes after `A` in the corpus, or just a random sentence? For example:

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Training with Cloud TPUs

Everything that we've described so far might seem fairly straightforward, so what's the missing piece that made it work so well? **Cloud TPUs**. Cloud TPUs gave us the freedom to quickly experiment, debug, and tweak our models, which was critical in allowing us to move beyond existing pre-training techniques. The **Transformer model architecture**, developed by researchers at Google in 2017, also gave us the foundation we needed to make BERT successful. The Transformer is implemented in our **open source release**, as well as the **tensor2tensor library**.

Results with BERT

To evaluate performance, we compared BERT to other state-of-the-art NLP systems. Importantly, BERT achieved all of its results with almost no task-specific changes to the neural network architecture. On **SQuAD v1.1**, BERT achieves 93.2% F1 score (a measure of accuracy), surpassing the previous state-of-the-art score of 91.6% and human-level score of 91.2%:

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

BERT also improves the state-of-the-art by 7.6% absolute on the very challenging GLUE benchmark, a set of 9 diverse Natural Language Understanding (NLU) tasks. The amount of human-labeled training data in these tasks ranges from 2,500 examples to 400,000 examples, and BERT substantially improves upon the state-of-the-art accuracy on all of them:

Rank	Model	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	QNLI	RTE
1	BERT: 24-layers, 1024-hidden, 16-heads	80.4	60.5	94.9	85.4/89.3	87.6/86.5	89.3/72.1	86.7	91.1	70.1
2	Singletask Pretrain Transformer	72.8	45.4	91.3	75.7/82.3	82.0/80.0	88.5/70.3	82.1	88.1	56.0
3	BiLSTM+ELMo+Attn	70.5	36.0	90.4	77.9/84.9	75.1/73.3	84.7/64.8	76.4	79.9	56.8

Making BERT Work for You

The models that we are releasing can be fine-tuned on a wide variety of NLP tasks in a few hours or less. The open source release also includes code to run pre-training, although we believe the majority of NLP researchers who use BERT will never need to pre-train their own models from scratch. The BERT models that we are releasing today are English-only, but we hope to release models which have been pre-trained on a variety of languages in the near future.

through Colab with the notebook "BERT Fine-tuning with Cloud TPUs".

You can also read our paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" for more details.

Labels: Deep Learning Natural Language Processing Natural Language Understanding open source TensorFlow

Previous posts

OCT 31, 2018

Google at EMNLP 2018

OCT 30, 2018

Introducing AdaNet: Fast and Flexible AutoML with Learning

OCT 29, 2018

Acoustic Detection of Humpback Whales Using a Convolutional

PX0392

Snapchat Support

What can we help you with?

Snapchat Support > Using Snapchat > My Friends > Adding and Removing Friends

How to Block a Friend on Snapchat

When you block a friend, they won't be able to view your Story or Charms, or send you Snaps or Chats.

To block a friend...

1. Swipe right to go to the **Chat screen**
2. Tap and hold on a friend's name
3. Tap '**Manage Friendship**'
4. Tap '**Block**'

Or...

1. Swipe right to go to the **Chat screen**
2. Tap on a **Chat** with that friend
3. Tap their profile icon at the top
4. Tap the ••• at the top, then tap '**Manage Friendship**'
5. Tap '**Block**'

Was this article helpful?

Related articles

How to Unblock a Friend on Snapchat

How to Remove a Friend on Snapchat

How do I change my privacy settings on Snapchat?

Troubleshoot Issues Adding Friends on Snapchat

Snapchat Support

Company	Community	Advertising	Legal
Snap Inc.	Snapchat Support	Snapchat Ads	Snap Terms
Careers	Pixy Support	Advertising Policies	Law Enforcement
News	Community Guidelines	Political Ads Library	Cookie Policy
Privacy and Safety		Brand Guidelines	Cookie Settings
		Promotions Rules	Report Infringement

Snap Inc.

Privacy Policy

Terms of Service

Language English (US)

PX0393

Snapchat Support

What can we help you with?

Snapchat Support > Using Snapchat > My Friends > Adding and Removing Friends

How to Remove a Friend on Snapchat

When you remove a friend from your friends list, they won't be able to view any of your private Stories or Charms, but they'll still be able to view any content you have set to public. Depending on your privacy settings, they may also still be able to Chat or Snap you!

To remove a friend from your friends list...

1. Swipe right to go to the **Chat screen**
2. Tap and hold on a friend's name
3. Tap '**Manage Friendship**'
4. Tap '**Remove Friend**'

Or...

1. Swipe right to go to the **Chat screen**
2. Tap on a friend's profile icon to open their Profile
3. Tap the ••• at the top, then tap '**Manage Friendship**'
4. Tap '**Remove Friend**'

Please Note: removing, blocking, or muting a friend should remove them from the Stories screen.

Was this article helpful?

Related articles

How do I deactivate or delete my Snapchat account?

Troubleshoot Issues Adding Friends on Snapchat

How do I change my privacy settings on Snapchat?

Snapchat Support

How to Block a Friend on Snapchat

Company

Snap Inc.

Careers

News

Privacy and Safety

Community

Snapchat Support

Pixy Support

Community Guidelines

Advertising

Snapchat Ads

Advertising Policies

Political Ads Library

Brand Guidelines

Promotions Rules

Legal

Snap Terms

Law Enforcement

Cookie Policy

Cookie Settings

Report Infringement

Snap Inc.

Privacy Policy

Terms of Service

Language English (US)

PX0394

Snapchat Support

What can we help you with?

Snapchat Support > Privacy > Privacy Settings

How do I change my privacy settings on Snapchat?

By default, only 'Friends' you've added on Snapchat can contact you directly or view your Story.

How to change your privacy settings

1. Tap **the button** in the Profile screen to open **Settings**
2. Scroll down to the '**Privacy Controls**' section and tap an option
3. Choose an option, then tap the back button to save your choice

Privacy settings options

Who Can Contact Me: Choose who can contact you directly with Snaps, Chats, calls, etc.

Send Me Notifications: Choose to receive notifications from everyone or just your friends. Learn more about notifications settings for iOS and Android.

Who Can View My Story: Choose who can view your Story. Tap '**Custom**' if you'd like to block specific friends from seeing your Story.

If you are a Public Profile, you will have to go to your profile to adjust your My Story privacy settings.

Who Can See My Location: Choose who can view your location on the Snap Map. Your location won't be shared on the Map until you open it for the first time.

Who Can Use My Cameos Selfie: Choose who can use your Cameos selfie in two-person Cameos.

Who Can See Me In Quick Add: Choose who can see you in Quick Add, a feature that appears around Snapchat which makes it easier to add friends.

Activity Indicator: Choose to let other Snapchatters see if you have been active on Snapchat recently.

A few things to remember

Even if you choose '**My Friends**,' anyone you're in a Group with will be able to communicate with you in Group Chat. To see who's in a Group before you jump in, just press and hold on the name of the group in the Chat screen!

Snapchat Support

you Snaps and Chats

If you post a Snap to your Story, and then change your settings so only friends can see your Story, others may still be able to see the Snaps you posted before the change

Pro Tip If you'd like to receive Snaps from 'Everyone' but only want to receive notifications when 'Friends' Snap you, you can change your notification settings for iOS or Android in **Snapchat settings > Notifications**.

Was this article helpful?

Related articles

How do I deactivate or delete my Snapchat account?

I lost my Snapstreak. How do I restore it?

How do I control my Snapchat notifications on iOS?

My account is locked

How do I stay safe on Snapchat?

Recently viewed articles

How to Remove a Friend on Snapchat

How to Block a Friend on Snapchat

Snapchat Support

Snap Inc.

Privacy Policy

Terms of Service

Language English (US)

PX0395

Skip to main content

Help Center (https://help.twitter.com/en)

- Using X (https://help.twitter.com/en/using-x)
- Managing your account (https://help.twitter.com/en/managing-your-account)
- Safety and security (https://help.twitter.com/en/safety-and-security)
- Rules and policies (https://help.twitter.com/en/rules-and-policies)
- Resources
  - New user FAQ (https://help.twitter.com/en/resources/new-user-faq)
  - Glossary (https://help.twitter.com/en/resources/glossary)
  - A safer X (https://help.twitter.com/en/resources/a-safer-twitter)
  - Accessibility (https://help.twitter.com/en/resources/accessibility)
  - Our rules (https://help.twitter.com/en/resources/rules)
  - My privacy (https://help.twitter.com/en/resources/how-you-can-control-your-privacy)
  - How we address misinformation on X (https://communitynotes.twitter.com/guide/en/about/introduction)
  - Recommender Systems (https://help.twitter.com/en/resources/recommender-systems)

Contact Us (https://help.twitter.com/forms.html)

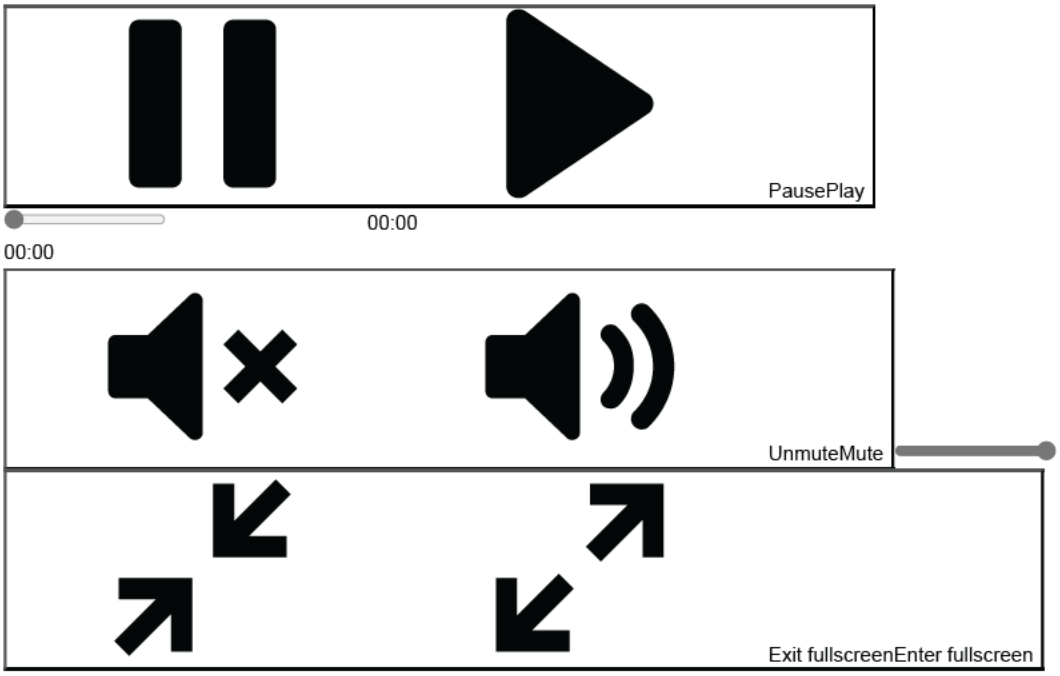
1. Help Center (https://help.twitter.com/en)
2. Blocking and muting (https://help.twitter.com/en/using-x#blocking-and-muting)
3. How to block accounts on X

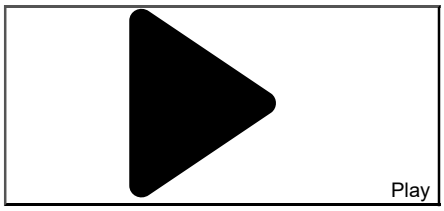
How to block accounts on X

1. Help Center (https://help.twitter.com/en)
2. Blocking and muting (https://help.twitter.com/en/using-x#blocking-and-muting)

How to block accounts on X

Block is a feature that helps you control how you interact with other accounts on X. This feature helps people in restricting specific accounts from contacting them, seeing their posts, and following them.





Note: Learn about our advanced block options for more detailed information on this feature.

Some important things to know about block:

- Accounts you have blocked cannot follow you, and you cannot follow an account you have blocked.
- Blocking an account you are currently following will cause you to unfollow that account (and them to unfollow you). If you decide to unblock that account, you will have to follow that account again.
- Blocked accounts do not receive a notification alerting them that their account has been blocked. However, if a blocked account visits the profile of an account that has blocked them, they will see they have been blocked (unlike mute, which is invisible to muted accounts).
- If you block an account and they choose to report your account, any of your posts that directly mention them will be available for them to view and attach during the reporting process.
- You will not receive notifications from accounts you block, or accounts that you do not follow who mention you in conversations started by accounts that you block. You will, however, see notifications from accounts you follow when they mention you in a conversation started by an account you block. If you'd like to view all of your mentions, you can do so by searching your username.

Important
Blocking only works if the account you've blocked is logged in on X. For example, if the account you've blocked isn't logged in or is accessing X content via a third party, they may be able to see your public posts. Please keep this in mind when you choose to share content on X.

Blocked accounts cannot:

- Follow you
- View your posts when logged in on X (unless they report you, and your posts mention them)
- Find your posts in search when logged in on X
- Send Direct Messages to you
- View your following or followers lists, likes or lists when logged in on X
- View a Moment you've created when logged in on X
- Add your X account to their lists
- Tag you in a photo

posts from blocked accounts will not appear in your timeline. However, please note that you may see posts or notifications in your timeline for the following:

1. posts from others you follow that mention accounts you have blocked.
2. posts that mention you, along with an account you have blocked.

Instructions for:

How to block a X account

To block from a post

Step 1

Tap the ••• icon located at the top of a post from the account you wish to block.

Step 2

Tap **Block**, and then select **Block** to confirm.

To block from a profile

Step 1

Visit the profile page of the account you wish to block.

Step 2

Tap the **more** icon •••

Step 3

Tap **Block**, and then select **Block** to confirm.

To block from a post

Step 1

Tap the ••• icon located at the top of a post from the account you wish to block.

Step 2

Tap **Block**, and then select **Block** to confirm.

To block from a profile

Step 1

Visit the profile page of the account you wish to block.

Step 2

Tap the overflow icon •••

Step 3

Tap **Block**, and then select **Block** to confirm.

To block from a post

Step 1

Click the ••• icon located at the top of a post from the account you wish to block.

Step 2
Click **Block**, and then select **Block** to confirm.
To block from a profile
Step 1
Go to the profile page of the account you wish to block.

Step 2
Click the more ••• icon on their profile page.

Step 3
Select **Block** from the menu.

Step 4
Click **Block** to confirm.

How do I know if I have blocked someone?

When you visit the profile of an account you have blocked, the **Follow** button will be replaced by a **Blocked** button.

The posts of an account you have blocked are hidden when you visit their profile. However, you can see the posts from that account by clicking the **Yes, view profile** button.

To unblock a X account:

1. Visit the blocked account's profile on X.
2. Click or tap the **Blocked** button.
3. Confirm you wish to unblock the account by selecting **Unblock** on X for iOS, and **Yes** on X for Android.

More resources

Sometimes you may find that blocking other accounts isn't the right solution—it either goes too far or doesn't go far enough toward changing your X experience. Find a complete list of other actions you can take to improve your X experience.

Additionally, explore our articles about staying safe and in control of your X experience (https://help.twitter.com/safety-and-security/), as well as our trusted resources (https://about.twitter.com/en_us/safety/safety-partners.html) article for a list of partners we work with to keep people safe across the world.

Share this article

Post

© 2024 X Corp.

Cookies (https://help.twitter.com/rules-and-policies/twitter-cookies)
Privacy (https://twitter.com/privacy)
Terms and conditions (https://twitter.com/tos)
English

Help Center (https://help.twitter.com/en)

- English (https://help.twitter.com/en/using-x/blocking-and-unblocking-accounts)
- Español (https://help.twitter.com/es/using-x/blocking-and-unblocking-accounts)
- 日本語 (https://help.twitter.com/ja/using-x/blocking-and-unblocking-accounts)
- 한국어 (https://help.twitter.com/ko/using-x/blocking-and-unblocking-accounts)
- Português (https://help.twitter.com/pt/using-x/blocking-and-unblocking-accounts)
- Deutsch (https://help.twitter.com/de/using-x/blocking-and-unblocking-accounts)
- Türkçe (https://help.twitter.com/tr/using-x/blocking-and-unblocking-accounts)
- Français (https://help.twitter.com/fr/using-x/blocking-and-unblocking-accounts)
- Italiano (https://help.twitter.com/it/using-x/blocking-and-unblocking-accounts)
- العربية (https://help.twitter.com/ar/using-x/blocking-and-unblocking-accounts)
- Nederlands (https://help.twitter.com/nl/using-x/blocking-and-unblocking-accounts)
- Bahasa Indonesia (https://help.twitter.com/id/using-x/blocking-and-unblocking-accounts)
- Русский (https://help.twitter.com/ru/using-x/blocking-and-unblocking-accounts)
- हिंदी (https://help.twitter.com/hi/using-x/blocking-and-unblocking-accounts)
- עברית (https://help.twitter.com/he)
- 简体中文 (https://help.twitter.com/zh-cn/using-x/blocking-and-unblocking-accounts)
- 繁體中文 (https://help.twitter.com/zh-tw/using-x/blocking-and-unblocking-accounts)
- ภาษาไทย (https://help.twitter.com/th/using-x/blocking-and-unblocking-accounts)
- Tiếng Việt (https://help.twitter.com/vi)
- Melayu (https://help.twitter.com/ms)
- Filipino (https://help.twitter.com/fil)
- فارسی (https://help.twitter.com/fa)
- Dansk (https://help.twitter.com/da)
- Suomi (https://help.twitter.com/fi)
- Svenska (https://help.twitter.com/sv)
- Norsk (https://help.twitter.com/no)
- Polski (https://help.twitter.com/pl)
- Magyar (https://help.twitter.com/hu)
- Română (https://help.twitter.com/ro)
- Українська (https://help.twitter.com/uk)
- Български (https://help.twitter.com/bg/using-x/blocking-and-unblocking-accounts)
- Català (https://help.twitter.com/ca/using-x/blocking-and-unblocking-accounts)
- Hrvatski (https://help.twitter.com/hr/using-x/blocking-and-unblocking-accounts)
- Српски (https://help.twitter.com/sr/using-x/blocking-and-unblocking-accounts)
- Slovenčina (https://help.twitter.com/sk/using-x/blocking-and-unblocking-accounts)
- پښتو (https://help.twitter.com/ps)
- Dari (https://help.twitter.com/fa-af)
- Oromo (https://help.twitter.com/om)
- Tigrinya (https://help.twitter.com/ti)
- Kurdish (https://help.twitter.com/ckb)
- Lietuvių (https://help.twitter.com/lt)
- Latviešu (https://help.twitter.com/lv)
- Malti (https://help.twitter.com/mt)
- Slovenščina (https://help.twitter.com/sl)
- Gaeilge (https://help.twitter.com/ga)
- Lus Hmoob (https://help.twitter.com/hmn)
- Հայերեն (https://help.twitter.com/hy)
- ខ្មែរ (https://help.twitter.com/km)

PX0396

Blocking users

Blocking users disables them from viewing your videos or engaging with you through direct messages, comments, follows or likes.

How to block someone:

1. Go to the person's profile.
2. Tap the **3-dot icon** in the top right.
3. Tap **Block** and follow the steps in the app.

How to unblock someone:

1. Go to the person's profile.
2. Tap the **3-dot icon** in the top right.
3. Tap **Unblock** and follow the steps in the app.

How to block people in bulk:

1. Long-press on a comment or tap the **pencil icon** in the upper left corner to open a window of options.
2. Tap **Manage multiple comments**.
3. Select up to 100 comments.
4. Tap **More**.
5. Tap **Block accounts**.

Using the bulk block feature, you can block up to 100 accounts at a time.

PX0397

User safety

Jump to a section

What is Restricted Mode? • How to turn Restricted Mode on or off • What is Comment Care Mode? • How to manage Comment Care Mode • What is filter video keywords? • What is Family Pairing? • How to set up Family Pairing • How to manage Family Pairing controls • Safety resources

At TikTok, we care about your safety while using the app. That's why we provide a number of safety tools and resources that allow you to control privacy preferences and access content that's most relevant or appropriate for you or your family.

What is Restricted Mode?

Restricted Mode on TikTok limits exposure to content that may not be suitable for everyone, for example, because it contains mature or complex themes. Some features will be unavailable under Restricted Mode, such as the Following feed and gifting on LIVE. Parents and guardians can also turn on Restricted Mode for their teens through Family Pairing.

How to turn Restricted Mode on or off

To turn Restricted Mode on or off:
1. In the TikTok app, tap **Profile** at the bottom.
2. Tap the **Menu** ☰ button at the top.
3. Tap **Settings and privacy.**
4. Tap **Content preferences**, then tap **Restricted Mode.**
5. Follow the steps in the app to set or enter a passcode to turn Restricted Mode on or off.

Note:
- This feature is only available on the TikTok app. When browsing on the web or mobile, Restricted Mode is on by default.
- If you have multiple accounts, you must turn on Restricted Mode on each account separately.

- If you're a parent or guardian, you may also turn on Restricted Mode for your teen's account through Family Pairing.

Learn more about Restricted Mode.

What is Comment Care Mode?

Comment Care Mode gives you more control over your experience by applying stronger filters to the comment areas of your new videos. Comment Care Mode filters comments that are inappropriate, offensive, and have been flagged by yourself or others.

How to manage Comment Care Mode

To turn on Comment Care Mode:
1. In the TikTok app, tap **Profile** at the bottom.
2. Tap the **Menu** ☰ button.
3. Tap **Settings and privacy**, then tap **Privacy**.
4. Tap **Comments**, then tap **Comment Care Mode**.
5. Turn on **Comment Care Mode**, then tap **Save**.

Learn more about Comment Care Mode.

What is filter video keywords?

Filter video keywords on TikTok allows you to tailor the content in your For You and Following feeds.

Note: Certain keywords can't be filtered.

To add video keyword filters:
1. In the TikTok app, tap **Profile** at the bottom.
2. Tap the **Menu** ☰ button at the top.
3. Tap **Settings and privacy.**
4. Tap **Content Preferences**, then tap **Filter video keywords**.
5. Tap **Add keywords** and enter a word or hashtag you'd like to filter. You can add up to 100 keywords.
6. Select the feeds you'd like to filter from.
7. Tap **Save** to confirm.

To filter hashtags directly from a video in the For You or Following feed:
1. In the TikTok app, press and hold on the video, or tap the **Share** button on the side, and tap **Not interested**.
2. Tap **Details** to display a list of hashtags found on the video.
3. Select the hashtags you'd like to filter.
4. Tap **Submit** to confirm.

Note: This option is not available on ads.

To delete or edit filtered keywords:
1. In the TikTok app, tap **Profile** at the bottom.
2. Tap the **Menu** ☰ button at the top.
3. Tap **Settings and privacy.**
4. Tap **Content Preferences**, then tap **Filter video keywords**.
5. Tap the **Delete** button next to any keywords you'd like to remove.
6. Tap **Delete** to confirm.

Note: If you're a parent or guardian, you can add keyword filters for your teen through Family Pairing.

What is Family Pairing?

Family Pairing on TikTok allows parents and teens to customize their safety settings based on individual needs. A parent can link their TikTok account to their teen's and set parental controls including:

Daily screen time

- Decide how long your teen can spend on TikTok each day. For teens between the ages 13 to 17, this setting is turned on by default to 1 hour.
- Set your teen's screen time limit directly from your own account.
- Set one screen time limit to apply to all your teen's devices.
- Set a passcode that you can choose to enter after your teen reaches their time to allow them to return to TikTok.

Screen time dashboard

- Get a summary of your teen's time spent on TikTok including:
  - The cumulative time spent each day for the last 4 weeks
  - The number of times your teen opened the app each day for the last 4 weeks

Learn more about screen time on TikTok.

Mute push notifications

- Decide when to mute your teen's push notifications. This setting is turned on by default:
  - For teens between the ages of 13 to 15 years, the scheduled time is set to 9:00 PM to 8:00 AM by default
  - For teens between the ages of 16 to 17 years, the scheduled time is set to 10:00 PM to 8:00 AM by default
- Schedule additional time directly from your own account. Keep in mind, the default scheduled time overrides any custom time that falls within the default time.

Learn more about how to mute push notifications on TikTok.

Filter video keywords

- Select keywords or hashtags to exclude from your teen's For You and Following feeds, as well as manage the visibility of your keywords list.

STEM feed

- Decide whether your teen can view the STEM feed.

Note: This is only available to users 13 years and older in the U.S.

Restricted Mode

- Restrict your teen's exposure to content that may not be appropriate or suitable for them.

Linked account activity

- Get notifications about your teen's activity, such as if your accounts get unlinked, by turning on **Customized updates and more** push notifications.

Search

- Decide whether your teen can search for videos, hashtags, or LIVE videos on TikTok.

Discoverability

- Decide whether your teen's account is private or public. With a private account, your teen can approve who can follow them and view their content.

Suggest account to others

- Decide whether your teen's account can be recommended to others. Learn more about suggested accounts on TikTok.

Direct Messages

- Restrict who can send messages to your teen, or turn off direct messaging completely.

Keep in mind:

- Direct messaging on TikTok is available only to registered account holders aged 16 and older.
- Direct messaging is automatically turned off for registered accounts between the ages of 13 and 15.

Learn more about direct messages on TikTok.

Liked videos

- Decide who can view your teen's liked videos.

Comments

- Decide who can comment on your teen's videos.

Note: TikTok's Family Pairing parental controls are available only on the TikTok mobile app and not on mobile and desktop browsers. If your teen accesses TikTok from a browser as well as the mobile app, you may want to use the parental controls available on the mobile and desktop browser, in addition to TikTok's mobile app parental controls.

To set up screen time limits and restrict your teen's TikTok usage on browsers, visit:

- Google's family center for Android and Chrome.
- Apple's parental controls for iPhones, iPads, Macs, and Safari.

How to set up Family Pairing

To link parent and teen accounts:
1. In the TikTok app, tap **Profile** at the bottom.
2. Tap the **Menu** ☰ button at the top.
3. Tap **Settings and privacy,** then tap **Family Pairing.**
4. Tap **Parent** or **Teen.**
5. Follow the steps in the app to link accounts.

How to manage Family Pairing controls

To manage or view Family Pairing controls:
1. In the TikTok app, tap **Profile** at the bottom.
2. Tap the **Menu** ☰ button at the top.
3. Tap **Settings and privacy,** then tap **Family Pairing.**
4. Select the account you want to manage.
5. Update the controls, as needed.

Safety resources

At TikTok, we want to ensure you have the safety resources you need to safeguard your online experience. The following resources are available if you or someone you know needs support.

Safety Center
Visit our Safety Center to learn more about safety resources and tools on TikTok. You can also find information on your account status and records of any reports you've submitted, and share feedback in the app version of Safety Center.

To visit our Safety Center in the TikTok app:
1. Tap **Profile** at the bottom of the screen.
2. Tap the **Menu** ☰ button at the top.
3. Tap **Settings and privacy,** then tap **Support**.
4. Tap **Safety Center**.

Global resources

- Family Online Safety Institute
- Internet Watch Foundation
- WePROTECT Global Alliance
- National Center for Missing & Exploited Children

Local resources

For a list of local organizations, visit our Safety Center and select your country or region.

Additional resources

You can find additional resources in our Safety Center for the following issues:

- Preventing child sexual abuse on TikTok
- Sexual Assault Resources
- Eating disorders
- Bullying prevention
- COVID-19
- Suicide and self-harm
- Digital well-being
- Parent and caregivers guide

PX0398

Skip to main content

Help Center (https://help.twitter.com/en)

- Using X (https://help.twitter.com/en/using-x)
- Managing your account (https://help.twitter.com/en/managing-your-account)
- Safety and security (https://help.twitter.com/en/safety-and-security)
- Rules and policies (https://help.twitter.com/en/rules-and-policies)
- Resources
  - New user FAQ (https://help.twitter.com/en/resources/new-user-faq)
  - Glossary (https://help.twitter.com/en/resources/glossary)
  - A safer X (https://help.twitter.com/en/resources/a-safer-twitter)
  - Accessibility (https://help.twitter.com/en/resources/accessibility)
  - Our rules (https://help.twitter.com/en/resources/rules)
  - My privacy (https://help.twitter.com/en/resources/how-you-can-control-your-privacy)
  - How we address misinformation on X (https://communitynotes.twitter.com/guide/en/about/introduction)
  - Recommender Systems (https://help.twitter.com/en/resources/recommender-systems)

Contact Us (https://help.twitter.com/forms.html)

1. Help Center (https://help.twitter.com/en)
2. Direct Messages (https://help.twitter.com/en/using-x#direct-messages)
3. About Direct Messages

About Direct Messages

1. Help Center (https://help.twitter.com/en)
2. Direct Messages (https://help.twitter.com/en/using-x#direct-messages)

About Direct Messages

As a Creator (https://help.twitter.com/using-x/subscriptions-creator#how) on 𝕏, as of Sept 28, 2023 you can now opt-in to allow messages from your subscribers. This feature is disabled by default, but you can enable it on the Messaging Settings page.

Direct Messages are the private side of X. You can use Direct Messages to have private conversations with people about posts and other content.

The basics

To send a Direct Message from X for iOS

To send a Direct Message from X for Android

To send a Direct Message via the web

To pin Direct Messages

To search Direct Messages

To snooze Direct Message notifications

To report a Direct Message or conversation

To share a post via Direct Message

To receive Direct Messages from anyone

To review Direct Message requests

To disable Direct Message read receipts

To send and receive Direct Messages on your phone via SMS

Some important things to know about Direct Messages

The basics

- You can start a private conversation or create a group conversation with anyone who follows you.
- Anyone you do not follow can send you a Direct Message if:
  - You have opted in to receive Direct Messages from anyone* or;
  - They are Verified and you have opted in to receive Direct Messages from Verified users or;
  - You have previously sent that person a Direct Message.

- Anyone in a conversation can send Direct Messages to the group. Everyone in a group can see all messages, even if everyone doesn't follow each other.
- In group conversations, anyone in the conversation can add other participants. Newly added participants won't see the prior history of the conversation.
- Some accounts, particularly businesses on X, have enabled a setting to receive Direct Messages from anyone. You can send a Direct Message to these accounts even if they don't follow you.
- Some accounts have enabled a setting to receive Direct Messages from Verified accounts. You can send a Direct Message to these accounts if your account is Verified.
- In both group and one-on-one conversations, you cannot be in a conversation with an account you block.

***Note**: For some people who create a new account on X, your account settings may already be set to receive message requests from other people you don't follow. These requests are kept separate from your other DMs until you accept them. You can accept the request to continue the conversation. To disable message requests, go to your Settings (https://twitter.com/settings/direct_messages).

To send a Direct Message from X for iOS

1. Tap the **envelope** icon. You'll be directed to your **messages**.
2. Tap the **message** icon to create a new message.
3. In the address box, enter the name(s) or @username(s) of the people you wish to send a message to. A group message can include up to 256 people.
4. Enter your message.
5. In addition to text, you can include a photo, video, or GIF with your Direct Message. From the message compose bar, or the **plus** icon you can access the following options:
   - Tap the **photo** icon to take a photo or record a video, or attach one from your device gallery. You have the option to edit your photo before sending your message. Tap on the photo to bring up editing screen where you can enhance, crop, and add filters. When you are done editing, tap **Save**. Learn more about advanced photo options.
   - To include an animated GIF in your message, tap the **GIF** icon to search for and select a file from the media library.
6. To send your message, tap the **paper airplane** icon

To delete a Direct Message or conversation:

- To delete a Direct Message, tap and hold the message and select **Delete message** from the menu that pops up.
- To delete an entire conversation from your inbox, swipe left on the conversation and tap the **trash can** icon You can also delete an entire conversation by tapping the **information** icon and selecting **Delete conversation** from the **Conversation info** page.
- When you delete a Direct Message or conversation (sent or received), it is deleted from your account only. Others in the conversation will still be able to see Direct Messages or conversations that you have deleted.

To manage a group conversation:

- To quickly access the list of conversation participants, tap the **profile** photo of the group conversation from your inbox.
- From within a group conversation, tap on the **information** icon ⓘ to bring up a settings page.
- Settings details:
  - Tap **Edit** to update the group conversation profile photo and name. To change the photo, tap the **camera** icon in the photo to **Choose from library** or **Take new photo**. Tap **Save** to update.
    Note: Once you update the photo, you'll have the option to **Remove current photo**, **View current photo**, **Choose from library**, or **Take new photo**.
  - Tap **Add members** to add people to the conversation.
  - Drag the slider next to **Snooze conversation** to snooze notifications for **1 hour**, **8 hours**, **1 week**, or **Forever**.
  - Drag the slider next to **Snooze mentions** to control whether you will receive notifications when you are mentioned in a group conversation. Please note that, unless this feature is enabled, you will receive notifications when you are directly mentioned in a conversation, even if you have enabled the **Snooze conversation** feature. Additionally, you must be a participant in a group conversation to receive mention notifications for that conversation.
  - To report the group conversation, tap **Report conversation**.
  - To remove yourself from the group conversation, tap **Leave conversation**.

To send a Direct Message from X for Android

1. Tap the **envelope** icon. You'll be directed to your messages.
2. Tap the **message** icon to create a new message.
3. In the address box, enter the name(s) or @username(s) of the people you wish to send a message to. A group message can include up to 256 people.
4. Enter your message.
5. In addition to text, you can include a photo, video, or GIF via Direct Message.
    - Tap the **photo** icon to take a photo or record a video, or attach one from your device gallery. You have the option to edit your photo through your X for iOS or X for Android app before sending your message. Tap on the photo to bring up editing screen where you can enhance, crop, and add filters. When you are done editing, tap **Save**. Learn more about advanced photo options.
    - To include an animated GIF in your message, tap the **GIF** icon to search for and select a file from the media library.
6. Tap the **send** icon.

To delete a Direct Message or conversation:

- To delete a Direct Message, tap and hold the message and select **Delete message** from the menu that pops up.
- To delete an entire conversation from your inbox, tap and hold the conversation and select **Delete conversation**. You can also delete an entire conversation by tapping the **information** icon and selecting **Delete conversation** from the **Conversation info** page.
- When you delete a Direct Message or conversation (sent or received), it is deleted from your account only. Others in the conversation will still be able to see Direct Messages or conversations that you have deleted.

To manage a group conversation:

- To quickly access the list of conversation participants, tap the profile photo of the group conversation from your inbox.
- From within a group conversation, tap on the **information** icon ⓘ to bring up a settings page.
- Settings details:
  - Tap **Edit** to update the group conversation profile photo and name. Tap the **camera** icon to bring up the following photo options: **View photo**, **Camera**, **Photo Gallery**, or **Remove photo**. Tap **Save** to update.
  - Tap **Add members** to add people to the conversation. The creator of the group is the default admin. If the creator is no longer in the group, the first member to join the group after the admin will become the admin. As the group admin, you are able to remove members from the group.
  - Tap **Snooze conversation** to snooze notifications for **1 hour**, **8 hours**, **1 week**, or **Forever**.
  - Check the box next to **Snooze mentions** to control whether you will receive notifications when you are mentioned in a group conversation. Please note that, unless this feature is enabled, you will receive notifications when you are directly mentioned in a conversation, even if you have enabled the **Snooze conversation** feature. Additionally, you must be a participant in a group conversation to receive mention notifications for that conversation.
  - To report the group conversation, tap **Report conversation**.
  - To remove yourself from the group conversation, tap **Leave conversation**.

To send a Direct Message via the web

1. Click **Messages** in the left navigation bar.
2. You'll see your Direct Message history. Click the **New message** icon at the top.
3. In the address box, enter the name(s) or @username(s) of the people you wish to send a message to. A group message can include up to 256 people.
4. Click **Next**
5. In the message box, you can include a photo, video, GIF, or emoji via Direct Message:

- Click the photo icon to upload a photo or video.
- To include an animated GIF in your message, tap the GIF icon to search for and select a file from the media library.
- Click the **Send** button or press the enter key to send.

Note: To start a new line in a message, press the shift and enter keys at the same time. Pressing only the enter key will send your message.

To delete a Direct Message or conversation:

- To delete a Direct Message, click on the message and select **Delete**.
- To leave a conversation, locate the conversation and click to open it. Click the **information** ⓘ icon, and select **Leave conversation**.
- When you delete a Direct Message or leave a conversation (sent or received), it is deleted from your account only. Others in the conversation will still be able to see Direct Messages or conversations that you have deleted.

To manage a group conversation:

- From within a group conversation, click on the **information** icon ⓘ to access the conversation settings:
  - From within the **Group info** page, click the **more** icon ••• to access the drop-down menu. You can choose to **Edit group name**, **Upload new photo**, **View photo,** or **Remove photo**.
    Note: The option to view and remove the photo will be available only if a group message photo is uploaded.

- Under **Notifications**, you can select to:
  - Click **Snooze notifications** to snooze notifications for **1 hour**, **8 hours**, **1 week**, or **Forever**.
  - Click **Snooze mentions** to control whether you will receive notifications when you are mentioned in a group conversation. Please note that, unless this feature is enabled, you will receive notifications when you are directly mentioned in a conversation, even if you have enabled the **Snooze conversation** feature. Additionally, you must be a participant in a group conversation to receive mention notifications for that conversation.

- To report the group conversation, click **Report conversation**.
- To remove yourself from the group conversation, click **Leave conversation**.

To search Direct Messages

Revisit conversations by searching keywords and names in the Direct Messages search bar.

How to search your Direct Messages
Navigate to your DMs and tap or click on the search bar. From there, type in keywords and names of the conversations you're looking for and messages, groups, and people with those keywords or names will populate. To go directly to the message, tap or click on the search result.

Note: Currently, searching images or GIFs to find DM conversations is not available.

To pin Direct Messages

This feature lets you pin your favorite conversations to the top of your DM inbox so you can find them easily. Pin up to six conversations at a time.

Instructions for:

How to pin a conversation

Step 1

Navigate to your **Direct Message inbox**.

Step 2

Swipe from left to right on the conversation you want to keep at the top of your DM inbox.

Step 3

Tap the **pin** icon.

Your pinned message will now appear at the top of your inbox. Pinned conversations appear from top to bottom in the order that you pin them.

To remove a conversation from your Pinned Conversations list, swipe left to right and tap the **pin** icon again.

Step 1

Navigate to your **Direct Message inbox**.

Step 2

Hold down on the conversation you want to keep at the top of your DM inbox.

Step 3

Tap the **pin** icon.

Your pinned message will now appear at the top of your inbox. Pinned conversations appear from top to bottom in the order that you pin them.

To remove a conversation from your Pinned Conversations list, hold down and tap the **pin** icon again.

Step 1
Navigate to your **Direct Message inbox**.

Step 2
Go to the conversation you want to keep at the top of your DM inbox and tap or click on the **more** icon.

Step 3
Select **Pin Conversation**.

Your pinned message will now appear at the top of your inbox. Pinned conversations appear from top to bottom in the order that you pin them.

To remove a conversation from your Pinned Conversations list, tap or click on the **more** icon to unpin the conversation.

To snooze Direct Message notifications

You can snooze notifications for Direct Messages for **1 hour**, **8 hours**, **1 week**, or **Forever**. When you snooze a Direct Message conversation, you will still receive new messages, but you won't receive a notification each time. **Note:** Unless the **Snooze mentions** setting is enabled, you will still receive notifications when you are directly mentioned in a group conversation that you are a part of.

How to snooze from within a Direct Message conversation on X.com, X for iOS, and X for Android:

1. Navigate to the Direct Message you'd like to snooze.
2. Click or tap into the message settings.
3. Click or tap the **information** icon ⓘ then select **Snooze notifications**.
4. From the pop-up menu, select the snooze time interval you'd like: **1 hour**, **8 hours**, **1 week**, or **Forever**.
5. To unsnooze, click or tap the **information** icon, then click or tap the slider for **snooze notifications**.

How to snooze from your Direct Message inbox on X for iOS:

1. Navigate to your Direct Message inbox.
2. Find the conversation you'd like to snooze.
3. Swipe left on the message and tap the **notifications** icon

   Note: You can also choose to **report** or **delete** a conversation when you swipe left.
4. From the pop-up menu, select the snooze time interval you'd like: **1 hour**, **8 hours**, **1 week**, or **Forever**.
5. To unsnooze, swipe left and tap the **snoozed notifications** icon

How to snooze from your Direct Message inbox on X for Android:

1. Navigate to your Direct Message inbox.
2. Find the conversation you'd like to snooze.
3. Long-press on the message and tap **Snooze notifications**.
4. From the pop-up menu, select the snooze time interval you'd like: **1 hour**, **8 hours**, **1 week**, or **Forever**.
5. To unsnooze, long-press on the message and tap **Unsnooze notifications**.

How to snooze from a Direct Message push notification:

1. Find the Direct Message push notification on your mobile device's lock screen you'd like to snooze:
    1. From the X for iOS app: Swipe left on the lock screen push notification, tap **View**, then tap **Snooze for 1 hour**.
    2. From the X for Android app: Swipe down on the lock screen push notification, then tap **Snooze**.

2. Conversation notifications will be snoozed for one hour.

To report a Direct Message or conversation

You can report an individual message or an entire conversation. Learn how to report a post or Direct Message for violations.

To share a post via Direct Message

Sharing a post via Direct Message is a great way to spark a conversation with a group of friends.

To receive Direct Messages from anyone

You can receive messages from anyone if you check the box next to **Allow Messages requests from everyone** in your Privacy and safety settings (https://twitter.com/settings/safety) on twitter.com. You can also adjust this setting via the X for iOS or X for Android apps. If you enable this option, anyone can message you and add you to group conversations.

If you're a Creator (https://help.twitter.com/using-twitter/subscriptions-creator#how) on 𝕏, you can select the checkbox next to **Allow messages from my subscribers** which will allow your subscribers to send you messages independent of other settings you can select. This setting is off by default.

- If you're a Subscriber with an active Subscription, you will be able to message the Creator if the setting is enabled. However, the Creator can disable this setting at any time.

- If you're a Creator who has enabled this setting and has replied to a message from your subscriber, they will be able to continue messaging you even if you subsequently disable the setting. If you don't reply to the initial message and then disable the setting, your subscriber won't be able to send you messages.

To change your settings using X for iOS:

1. Tap the **navigation menu**.
2. Select **Settings and privacy**.
3. Tap **Privacy and safety**.
4. Under **Direct Messages (https://twitter.com/settings/direct_messages)**, and next to **Allow messages requests from everyone**, drag the slider to allow anyone to send you Direct Messages.

To change your settings using X for Android:

1. In the top menu, you will either see a **navigation menu** icon ☰ or your **profile** icon. Tap whichever icon you have.
2. Select **Settings and privacy**.
3. Tap **Privacy and safety**.
4. Under **Direct Messages**, and next to **Receive messages from anyone**, check the box to allow anyone to send you Direct Messages.

To change your settings using X for twitter.com:

1. Click on the **More** ⋯ icon in the navigation bar.
2. Select **Settings and privacy**.
3. Tap **Privacy and safety**.
4. Under **Direct Messages**, and next to **Receive messages from anyone**, check the box to allow anyone to send you Direct Messages.

Note: Disabling the **Receive messages from anyone** setting will not prevent you from continuing to receive Direct Messages from someone you don't follow if you have a prior conversation already established with that person. You will need to either **report** the conversation or **block** the account to stop receiving Direct Messages from that person.

To review Direct Message requests

If you have the **Receive messages from anyone** setting enabled, incoming messages from people you don’t follow will appear as **Requests** in the **Messages** tab. New group conversations that you’re added to by people you don’t follow will also appear in **Requests**. Entering the conversation, you will be asked to either **Delete** or **Accept** the message. Accepting the message will allow you to engage with the person, and will move the message to your inbox. Please note that they will not know you have seen the message until you have accepted their request.

Deleting the message will remove it from your inbox. **Note**: Deleting a message will not prevent that account from sending you messages in the future. You will always have the option to block the account or report the conversation. Blocked accounts cannot send you messages, unless you unblock them.

Accepting the message will allow you to engage with the person. All media will be hidden prior to accepting the message. If you wish to view the hidden media, click or tap **View media**.

Note: The option to **Accept** or **Delete** messages and **View Media** from someone new that you don't follow is only available on the X for iOS and Android apps, and X.com.

Additionally, by default, we filter lower-quality requests from the **Requests** section of your inbox for the X for iOS and Android app. When enabled, the quality filter for message requests hides conversation requests we think may be lower quality. You

will not receive notifications for filtered requests, but these messages will still be viewable behind the low quality filter that exists at the bottom of the **Requests** section of your inbox.

To disable or enable the quality filter for message requests:

- At the top of your **Requests** tab, tap **Change settings**.
- In the **Privacy** section of your **Message settings**, toggle the slider next to **Quality filter**.

How to filter graphic media in Direct Messages

The filter is enabled by default and will work in this way:

- There will be a **displayed warning** over graphic media in DMs, for both people you follow and unknown senders. Additionally, if it's from someone you don't follow, we'll also treat it like spam and move it to the bottom of your **Request** inbox.
- If turned off, and you receive a graphic message from someone you do follow, we will show the graphic media in the same manner as any other image, video, or GIF in the conversation you may receive. If it's from someone you don't follow, we'll still display a warning because we do that with all media, but it won't explicitly identify the media as potentially sensitive/graphic.

Note: If you have the **Allow message requests from everyone** setting enabled under the **Direct Messages** section of your **Privacy and safety** settings, it is possible to disable and enable the quality filter from your privacy settings. This feature is only available currently on iOS and Android devices.

To disable Direct Message read receipts

Direct Messages feature read receipts so you know when people have seen your messages. When someone sends you a Direct Message and your **Show read receipts** setting is enabled, everyone in the conversation will know when you've seen it. This setting is enabled by default but you can turn it off (or back on) through your settings at any time. If you turn off the **Show read receipts** setting, you will not be able to see read receipts from other people.

Read receipts are only viewable on the X for iOS and Android apps, and X.com. Please note, however, that read receipts will also be sent when you view Direct Messages on mobile web.

To disable or enable using X for iOS:

1. In the top menu, tap your **profile** icon.
2. Select **Settings and privacy**.
3. Tap **Privacy and safety.**
4. Under **Direct Messages**, and next to **Show read receipts**, drag the slider to turn the feature off or back on.

To disable or enable using X for Android:

1. In the top menu, you will either see a **navigation menu** icon ☰ or your **profile** icon. Tap whichever icon you have.
2. Select **Settings and privacy**.
3. Tap **Privacy and safety**.
4. Under **Direct Messages**, and next to **Show read receipts**, uncheck the box to turn the feature off or check the box to turn back on.

To disable or enable via X.com:

1. Click the **more** ⋯ icon, then click **Privacy and safety**.
2. Under **Privacy and safety** and next to **Show read receipts**, click the box to turn the feature off or click the box to turn back on.

Note: For conversations that appear under **Requests**, the sender(s) will not be able to see if you have read their Direct Messages unless you accept the conversation.

To send and receive Direct Messages on your phone via SMS

If your X account is connected to your mobile phone, you can send and receive Direct Messages via SMS.

- Haven't added your mobile phone yet? Find out how to add your phone number to your account.
- Added your mobile phone already? Find out what text commands to use to send Direct Messages via SMS.
- Read this article about setting up your mobile preferences to learn more about how to receive Direct Messages via SMS.

Note: **A note about Direct Message failures: Please be careful to ensure that Direct Messages you send via SMS are under 160 characters, including the d command and username.** Your service provider may split SMS messages greater than 160 characters into multiple SMS messages. In this case, the second and any subsequent SMS messages will post as public posts because they don't begin with the appropriate text command (**d username**) that tells X they are Direct Messages, as the first SMS message did.

To receive Direct Messages from Verified users

- You can receive messages from Verified users if "Allow messages requests from Verified users and people you follow only" is enabled in your Direct Messages settings (https://twitter.com/settings/safety), which you can check from our website (https://twitter.com/settings/safety), or our iOS and Android apps. When this option is enabled, Verified users can message you and add you to group conversations. If you are not currently following these users, these messages will appear in your message requests inbox.

Some important things to know about Direct Messages

- When you delete a Direct Message or conversation (sent or received), it is deleted from your account only. Others in the conversation will still be able to see Direct Messages or conversations that you have deleted. When you delete a group conversation, you will leave that group and will no longer be able to participate.
- When you share a link in a Direct Message, it is automatically processed and shortened to a t.co link. Learn more about link shortening. Please note that anyone with a t.co shortened link will be able to navigate to the destination URL.
- When you share media in a Direct Message, it will be viewable by everyone in the conversation. Please note that recipients may download or re-share links to media that you share in a Direct Message. Anyone with the link to media shared in a Direct Message will be able to view the content.

Need more information?

Check out the most frequently asked questions about Direct Messages.

Share this article

Post

© 2024 X Corp.

Cookies (https://help.twitter.com/rules-and-policies/twitter-cookies)

Privacy (https://twitter.com/privacy)

Terms and conditions (https://twitter.com/tos)

English

Help Center (https://help.twitter.com/en)

- English (https://help.twitter.com/en/using-x/direct-messages)
- Español (https://help.twitter.com/es/using-x/direct-messages)
- 日本語 (https://help.twitter.com/ja/using-x/direct-messages)
- 한국어 (https://help.twitter.com/ko/using-x/direct-messages)
- Português (https://help.twitter.com/pt/using-x/direct-messages)
- Deutsch (https://help.twitter.com/de/using-x/direct-messages)
- Türkçe (https://help.twitter.com/tr/using-x/direct-messages)
- Français (https://help.twitter.com/fr/using-x/direct-messages)
- Italiano (https://help.twitter.com/it/using-x/direct-messages)
- العربية (https://help.twitter.com/ar/using-x/direct-messages)
- Nederlands (https://help.twitter.com/nl/using-x/direct-messages)
- Bahasa Indonesia (https://help.twitter.com/id/using-x/direct-messages)
- Русский (https://help.twitter.com/ru/using-x/direct-messages)
- हिंदी (https://help.twitter.com/hi/using-x/direct-messages)
- עברית (https://help.twitter.com/he)
- 简体中文 (https://help.twitter.com/zh-cn)
- 繁體中文 (https://help.twitter.com/zh-tw)
- ภาษาไทย (https://help.twitter.com/th)
- Tiếng Việt (https://help.twitter.com/vi)
- Melayu (https://help.twitter.com/ms)
- Filipino (https://help.twitter.com/fil)
- فارسی (https://help.twitter.com/fa)
- Dansk (https://help.twitter.com/da)
- Suomi (https://help.twitter.com/fi)
- Svenska (https://help.twitter.com/sv)
- Norsk (https://help.twitter.com/no)
- Polski (https://help.twitter.com/pl)
- Magyar (https://help.twitter.com/hu)
- Română (https://help.twitter.com/ro)
- Українська (https://help.twitter.com/uk)
- Български (https://help.twitter.com/bg)
- Català (https://help.twitter.com/ca)
- Hrvatski (https://help.twitter.com/hr)
- Српски (https://help.twitter.com/sr)
- Slovenčina (https://help.twitter.com/sk)
- پښتو (https://help.twitter.com/ps)
- Dari (https://help.twitter.com/fa-af)
- Oromo (https://help.twitter.com/om)
- Tigrinya (https://help.twitter.com/ti)
- Kurdish (https://help.twitter.com/ckb)

- Lietuvių (https://help.twitter.com/lt)
- Latviešu (https://help.twitter.com/lv)
- Malti (https://help.twitter.com/mt)
- Slovenščina (https://help.twitter.com/sl)
- Gaeilge (https://help.twitter.com/ga)
- Lus Hmoob (https://help.twitter.com/hmn)
- Հայերեն (https://help.twitter.com/hy)
- ខ្មែរ (https://help.twitter.com/km)

PX0399

Skip to main content

Help Center (https://help.twitter.com/en)

- Using X (https://help.twitter.com/en/using-x)
- Managing your account (https://help.twitter.com/en/managing-your-account)
- Safety and security (https://help.twitter.com/en/safety-and-security)
- Rules and policies (https://help.twitter.com/en/rules-and-policies)
- Resources
  - New user FAQ (https://help.twitter.com/en/resources/new-user-faq)
  - Glossary (https://help.twitter.com/en/resources/glossary)
  - A safer X (https://help.twitter.com/en/resources/a-safer-twitter)
  - Accessibility (https://help.twitter.com/en/resources/accessibility)
  - Our rules (https://help.twitter.com/en/resources/rules)
  - My privacy (https://help.twitter.com/en/resources/how-you-can-control-your-privacy)
  - How we address misinformation on X (https://communitynotes.twitter.com/guide/en/about/introduction)
  - Recommender Systems (https://help.twitter.com/en/resources/recommender-systems)

Contact Us (https://help.twitter.com/forms.html)

1. Help Center (https://help.twitter.com/en)
2. Platform Use Guidelines (https://help.twitter.com/en/rules-and-policies#platform-use-guidelines)
3. Our range of enforcement options

Our range of enforcement options

1. Help Center (https://help.twitter.com/en)
2. Platform Use Guidelines (https://help.twitter.com/en/rules-and-policies#platform-use-guidelines)

Our range of enforcement options

When we take enforcement actions, we may do so either on a specific piece of content (e.g., an individual post or Direct Message), on an account, or employ a combination of these options. In some instances, this is because the behavior violates the X Rules. Other times, it may be in response to a valid legal request from an authorized entity in a given country. Below are some of the enforcement actions that we may take.

Post-level enforcement

We take action at the post level when a specific post violates the X Rules, including posts that share or reproduce other posts by posting screenshots, quote-posting, or sharing post URLs that violate our Rules.

A few of the ways in which we might take action at the post level include:

Limiting post visibility: Where appropriate, we will restrict the reach of posts that violate our policies and create a negative experience for other users by making the post less discoverable on X. This can include:

- Excluding the post from search results, trends, and recommended notifications
- Removing the post from the For you and Following timelines
- Restricting the post's discoverability to the author's profile
- Downranking the post in replies
- Restricting Likes, replies, Reposts, Quote posts, bookmarks, share, pin to profile, or Edit post

Excluding the post from having ads adjacent to it: Starting in April 2023, posts identified as violating our Rules will begin to receive labels informing both post authors and viewers that we limited the post's visibility.

Authors will be able to submit an appeal on the label if they think we incorrectly limited their post's visibility.

Requiring post removal: When we determine that a post violated the X Rules and the violation is severe enough to warrant post removal, we will require the violator to remove it before they can post again. They will need to go through the process of removing the violating post or appealing our removal request if they believe we made an error. The post will be hidden from public view with a notice during this process.

Labeling a post: If we determine a post contains misleading or disputed information per our policies that could potentially lead to harm, we may add a label to the content to provide context and additional information to users. In these cases, Community Notes may also be visible on posts to provide additional context.

Notice of public interest exception: We may determine that it is in the public interest for a post that would otherwise be in violation of our rules to remain accessible on our service. When this occurs, we will place the post behind a notice and limit its visibility. Learn more about the public interest exception.

Direct Message-level enforcement

Stopping conversations between a reported violator and the reporter's account: In a private Direct Message conversation, when a participant reports the other account, we will stop the violator from sending messages to the account who reported them. The conversation will also be removed from the reporter's inbox. However, if the reporter decides to continue to send Direct Messages to the violator, the conversation will resume.

Placing a Direct Message behind a notice: In a group Direct Message conversation, the violating Direct Message may be placed behind a notice to ensure no one else in the group can see it again.

Account-level enforcement

We take action to suspend an account if we determine that a user has engaged in repeated violations of our policies and/or violated specific policies that cause significant risk to X (i.e. posting illegal content, attempts to manipulate our platform or spam users, using our platform to incite violence, etc.) or pose a threat to our users (fraud, user privacy violations, violent threats, targeted harassment, etc.).

Placing an account in read-only mode: We may temporarily limit an account's ability to post, Repost, or Like. The user can read their timelines and will only be able to send Direct Messages to their followers.

Verifying account ownership: We may require an account owner to verify ownership with a phone number or email address. Note that when an account has been locked pending completion of a challenge (such as being required to provide a phone number), it is removed from follower counts, Reposts, and Likes.

Users can appeal account suspensions if they believe we made an error. They can do this through the platform interface or by filing a report.

Actions we may take against non-violating content

Placing a post behind a notice: We may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through. Learn more about how to control whether you see sensitive media.

Withholding a post based on age: We restrict views of specific forms of sensitive media such as adult content for viewers who are under 18 or viewers who do not include a birth date on their profile with interstitials.

Withholding a post or account in a country: We may withhold access to certain content in a particular country if we receive a valid and properly scoped request from an authorized entity in that country. Read more about country withheld content.

© 2024 X Corp.

Cookies (https://help.twitter.com/rules-and-policies/twitter-cookies)
Privacy (https://twitter.com/privacy)
Terms and conditions (https://twitter.com/tos)

English

Help Center (https://help.twitter.com/en)

- English (https://help.twitter.com/en/rules-and-policies/enforcement-options)
- Español (https://help.twitter.com/es/rules-and-policies/enforcement-options)
- 日本語 (https://help.twitter.com/ja/rules-and-policies/enforcement-options)
- 한국어 (https://help.twitter.com/ko/rules-and-policies/enforcement-options)
- Português (https://help.twitter.com/pt/rules-and-policies/enforcement-options)
- Deutsch (https://help.twitter.com/de/rules-and-policies/enforcement-options)
- Türkçe (https://help.twitter.com/tr/rules-and-policies/enforcement-options)
- Français (https://help.twitter.com/fr/rules-and-policies/enforcement-options)
- Italiano (https://help.twitter.com/it/rules-and-policies/enforcement-options)
- العربية (https://help.twitter.com/ar/rules-and-policies/enforcement-options)
- Nederlands (https://help.twitter.com/nl/rules-and-policies/enforcement-options)
- Bahasa Indonesia (https://help.twitter.com/id/rules-and-policies/enforcement-options)
- Русский (https://help.twitter.com/ru/rules-and-policies/enforcement-options)
- हिंदी (https://help.twitter.com/hi)
- עברית (https://help.twitter.com/he)
- 简体中文 (https://help.twitter.com/zh-cn)
- 繁體中文 (https://help.twitter.com/zh-tw)
- ภาษาไทย (https://help.twitter.com/th)
- Tiếng Việt (https://help.twitter.com/vi)
- Melayu (https://help.twitter.com/ms)
- Filipino (https://help.twitter.com/fil)
- فارسی (https://help.twitter.com/fa)
- Dansk (https://help.twitter.com/da)
- Suomi (https://help.twitter.com/fi)
- Svenska (https://help.twitter.com/sv)
- Norsk (https://help.twitter.com/no)
- Polski (https://help.twitter.com/pl)
- Magyar (https://help.twitter.com/hu)
- Română (https://help.twitter.com/ro)
- Українська (https://help.twitter.com/uk)
- Български (https://help.twitter.com/bg)
- Català (https://help.twitter.com/ca)
- Hrvatski (https://help.twitter.com/hr)
- Српски (https://help.twitter.com/sr)
- Slovenčina (https://help.twitter.com/sk/rules-and-policies/enforcement-options)
- پښتو (https://help.twitter.com/ps)
- Dari (https://help.twitter.com/fa-af)
- Oromo (https://help.twitter.com/om)
- Tigrinya (https://help.twitter.com/ti)
- Kurdish (https://help.twitter.com/ckb)

- Lietuvių (https://help.twitter.com/lt)
- Latviešu (https://help.twitter.com/lv)
- Malti (https://help.twitter.com/mt)
- Slovenščina (https://help.twitter.com/sl)
- Gaeilge (https://help.twitter.com/ga)
- Lus Hmoob (https://help.twitter.com/hmn)
- Հայերեն (https://help.twitter.com/hy)
- ខ្មែរ (https://help.twitter.com/km)

PX0400

How We Prevent the Spread of False Information on Snapchat

September 8, 2022

With the midterm elections approaching in the United States, we wanted to highlight our longstanding approach to preventing the spread of false information on Snapchat, and the steps we continue to take to build on our strong foundation of preventing the spread of false information on our platform.

Our efforts have always started with the architecture of our platform. With Snapchat, we wanted to build something different to capture the spontaneity and fun of real-life conversations. From the beginning, we've built safety and privacy into the fundamental design of our platform. That's why Snapchat opens directly to a camera, instead of a feed of content, and is focused on connecting people who are already friends in real life. We have always wanted Snapchatters to be able to express themselves and have fun with their friends — without the pressure to grow a following, gain views, or earn likes. Snapchat reflects how we normally communicate face to face, or on the phone, because digital communication on Snapchat deletes by default. Across Snapchat, we limit the ability for unmoderated content to reach a large audience. We do this by holding amplified content to a higher standard to make sure that it complies with our content guidelines. While Snapchat has evolved over the years, we have always tried to build technology that enables creativity and prioritizes the safety, privacy, and well-being of our community.

In addition to our foundational architecture, there are a number of key policies that that help us to prevent the spread of false information on Snapchat:

- Our policies have long prohibited the spread of false information. Both our Community Guidelines, which apply equally to all Snapchatters, and our content guidelines, which apply to our Discover partners, prohibit spreading false information that can cause harm — including, for example, conspiracy theories, denying the existence of tragic events, unsubstantiated medical claims, or undermining the integrity of civic processes. This includes sharing media that is manipulated to be misleading about real-life events (including harmful deepfakes or shallow-fakes).
- Our approach to enforcing against content that includes false information is straightforward: we remove it. When we find content that violates our guidelines, our policy is to take it down, which immediately reduces the risk of it being shared more widely.
- Across our app, we don't allow unvetted content the opportunity to 'go viral.' Snapchat does not offer an open newsfeed where people or publishers can broadcast false information. Our Discover platform features content from vetted media publishers, and our Spotlight platform is proactively moderated before content is eligible to reach a large audience. We offer Group Chats, which are limited in size, are not recommended by algorithms, and are not publicly discoverable on our platform if you are not a member of that Group.
- We use human review processes to fact-check all political and advocacy ads. All political ads, including election-related ads and issue advocacy ads, must include a transparent "paid for" message that discloses the sponsoring organization, and we provide access to information about all ads that pass our review in our Political Ads Library. In connection with U.S. elections, we partner with the nonpartisan Poynter Institute to independently fact-check political ad statements. In addition, to help mitigate the risks of foreign interference in elections, we prohibit the purchase of political ads from outside the country in which the ad will run.
- We are committed to increasing transparency into our efforts to combat false information. Our most recent Transparency Report, which covered the second half of 2021, included several new elements, including data about our efforts to enforce against false information globally. During this period, we took action against 14,613 pieces of content and accounts for violations of our policies on false information — and we plan to provide more detailed breakdowns of these violations in our future reports.

To build on this, ahead of the midterm elections, we have also established dedicated internal processes for information-sharing and for monitoring the effectiveness of our policies and other harm mitigation efforts, ensuring we can calibrate our approach as needed. We are also engaging actively with researchers, NGOs, and other stakeholders from across the election integrity, democracy, and information integrity communities to ensure our safeguards are responsibly anchored in the wider context of emerging trends and informed by expert perspectives.

We're also focused on partnering with experts to promote greater information integrity. Through our Discover content platform, we focus on providing credible and accurate news coverage to our community, from publishers like The Wall Street Journal, The Washington Post, VICE, and NBC News.

We've also developed an expansive array of in-app resources to connect users with civic information, including about opportunities to register to vote, or even run for local office.

Doing our part to promote a responsible information environment remains a major priority across our company, and we will continue to explore innovative approaches to reach Snapchatters where they are, while strengthening our efforts to protect Snapchat from the risks of viral false information.

Back to News

Company

Snap Inc.

Careers

News

Privacy and Safety

Community

Snapchat Support

Pixy Support

Community Guidelines

Advertising

Snapchat Ads

Advertising Policies

Political Ads Library

Brand Guidelines

Promotions Rules

Legal

Snap Terms

Law Enforcement

Cookie Policy

Cookie Settings

Report Infringement

Snap Inc.

Privacy Policy

Terms of Service

Language English (US)

PX0401

Newsroom

Combating climate misinformation on Pinterest

April 8, 2022

Ensuring that Pinners find ideas from trusted sources no matter what type of inspiration they are looking to discover on the platform —from how they cook, the way they shop, build their home, and how to live a more sustainable life —is important to Pinterest. That's why today, Pinterest is rolling out a new climate misinformation policy to keep false and misleading claims around climate change off the platform. **Our new policy makes Pinterest the only major digital platform to have clearly defined guidelines against false or misleading climate change information, including conspiracy theories, across content and ads.**

As part of our Community guidelines on misinformation and disinformation, our climate misinformation policy removes content that may harm the public's well-being, safety or trust, including:

- Content that denies the existence or impacts of climate change, the human influence on climate change, or that climate change is backed by scientific consensus.
- False or misleading content about climate change solutions that contradict well-established scientific consensus.

Newsroom

- Harmful false or misleading content about public safety emergencies including natural disasters and extreme weather events.

All ads on Pinterest always have to comply with our Community guidelines. Additionally, we've updated our Advertising guidelines to explicitly prohibit any ads containing conspiracy theories, misinformation and disinformation related to climate change.

"Pinterest believes in cultivating a space that's trusted and truthful for those using our platform. This bold move is an expansion of our broader misinformation guidelines, which we first developed in 2017 to address public health misinformation, and have since updated to address new and emerging issues as they come to the forefront. The expanded climate misinformation policy is yet another step in Pinterest's journey to combat misinformation and create a safe space online," said Sarah Bromma, Pinterest's Head of Policy.

Searches for a greener life are rising on Pinterest. People are regularly turning to Pinterest to find ideas to incorporate sustainability into their entire lifestyle as searches for "zero waste tips" were 6X greater, "recycling clothes ideas" were 4X higher, "recycled home decor" increased by +95% and "zero waste lifestyle" increased by +64% compared to last year[1].

Tackling issues like climate change or misinformation are complex, and requires the support and collaboration of an entire ecosystem. We have partnered with experts including the Climate Disinformation Coalition and the Conscious Advertising Network to help inform and develop our policy based on common misinformation themes they're seeing across media platforms.

"Climate disinformation on digital platforms is a serious threat to the public support needed to solve the climate crisis. Pinterest has demonstrated great leadership by creating a community standard that includes a definition of climate misinformation, and we will continue to press all platforms for transparency and reporting on their actions. We encourage others to take note of Pinterest's efforts to reduce climate change disinformation," said Michael Khoo, Climate Disinformation Co-Chair at Friends of the Earth.

Inspire a better future

To fuel inspiration around living a greener life, Pinterest Creators from the US, UK, Australia, Germany, France, Japan, Brazil, Mexico, Argentina, Indonesia and India will launch a Creator Originals content series on the platform. Their Idea Pins will work to inspire Pinners to learn more about thrifting tips, upcycling clothes and minimizing food waste, among other topics. Additionally, Pinners searching for relevant topics like "eco friendly living" and "sustainable living" will see a Pin that will take them to an article of the day until Earth Day—featuring curated boards and inspired content around sustainability.

Newsroom

[redacted]

Green Girl Leah

We will continue to inspire a better future both on our platform and within our communities by supporting the work of Project Drawdown, Intersectional Environmentalist and Potential Energy Coalition, the organization powering Science Moms, with donations to support their programmatic work as well as ad credits to elevate their content on Pinterest. With our climate misinformation policy, Pinterest can help the world make more informed decisions in their everyday lives to create a better, greener future for all.

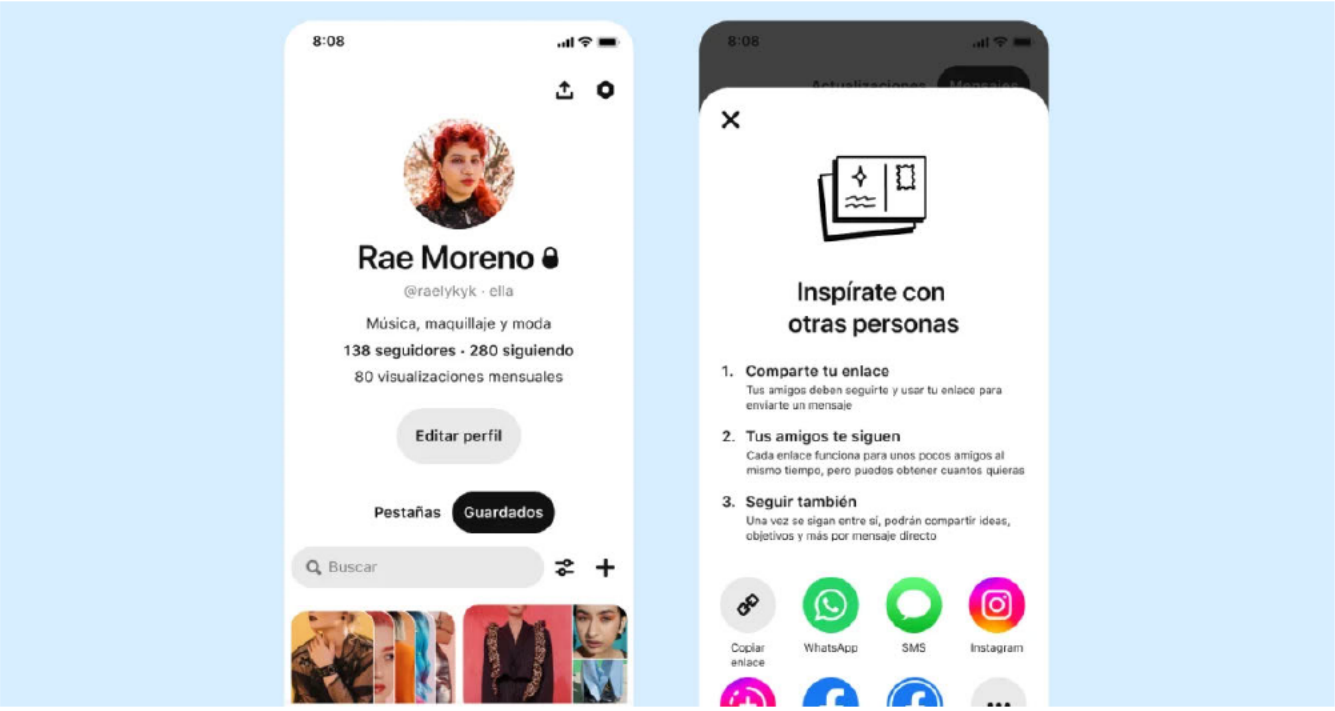
"It is unequivocal that climate misinformation combined with advertising tools delays meaningful climate action. The Conscious Advertising Network is delighted to partner with Pinterest who is taking a genuine leadership position by publishing this robust, well-considered and comprehensive climate misinformation policy. We look forward to the continued collaboration and call on all other tech platforms to follow Pinterest's lead," said Jake Dubbins, Co-Chair of Conscious Advertising Network.

Newsroom

September 13, 2023

Pinterest unveils slate of new product updates and ad solutions

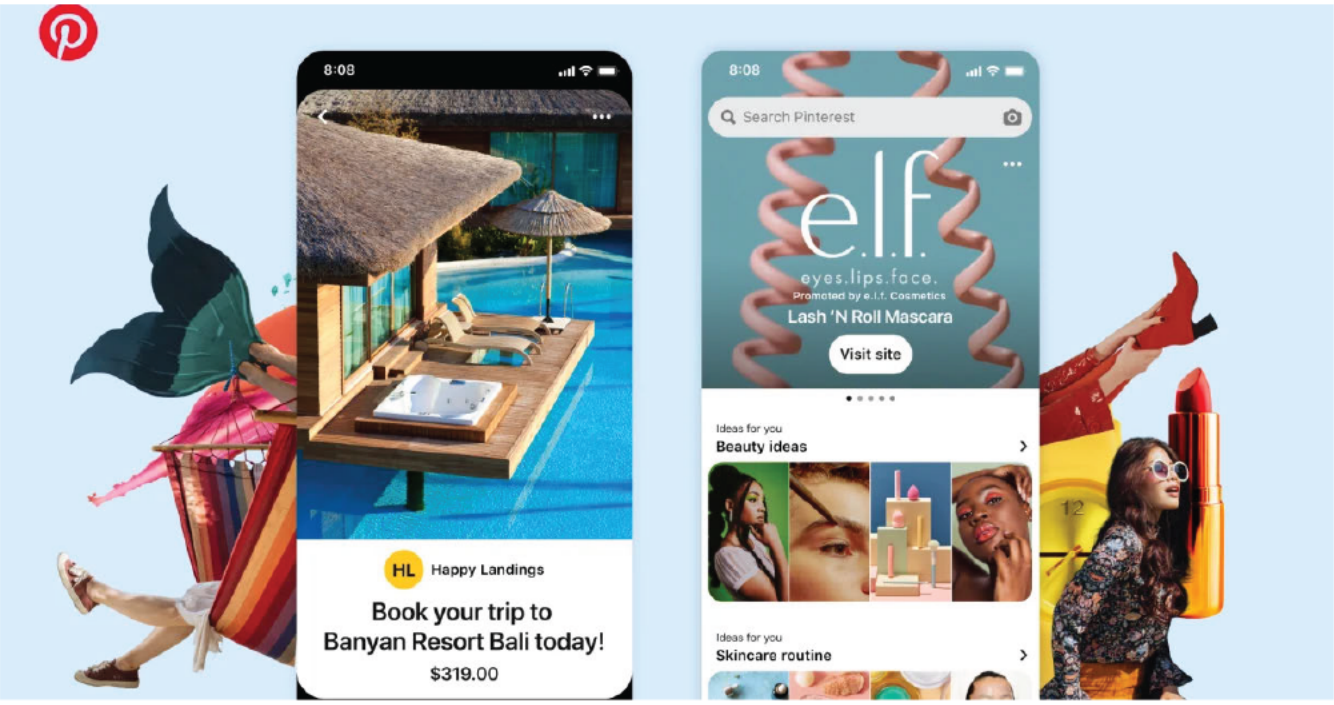
Newsroom



August 21, 2023

New features enhance teen safety on Pinterest

Newsroom



June 14, 2023

Introducing Premiere Spotlight and Travel Catalogs

Footnotes

English (US)

About

What's Pinterest

Our Pinterest profile

Contact

Resources

Help Center

Media gallery

Subscribe via RSS

Newsroom

For developers

For investors

Policies

Copyright & Trademark

Terms of service

Privacy & Cookies

Personalized ads

© 2024 Pinterest

PX0403

An update on our work to counter misinformation

Share this post

By Cormac Keenan, Head of Trust and Safety

TikTok empowers people to share their creativity, knowledge, and passion with others as they entertain and bring joy to our community. To help keep our platform welcoming and authentic for everyone, we remove content which violates our policies. We aim to be transparent about those removals with creators and through our Community Guidelines Enforcement Reports. Our next report for April-June, published today, shows improvements made in countering misinformation and we want to provide insight on the work that led to these gains.

Misinformation is not a new problem, but the internet provides a new avenue to an old challenge. We recognize the impact misinformation can have in eroding trust in public health, electoral processes, facts, and science. We are committed to being part of the solution. We treat misinformation with the utmost seriousness and take a multi-pronged approach to stopping it from spreading, while elevating authoritative information and investing in digital literacy education to help get ahead of the problem at scale.

Our policies

Our integrity policies aim to promote a trustworthy, authentic experience. Within those, our harmful misinformation policies prohibit content that could mislead our community about civic processes, public health, or safety. For instance, we do not allow medical misinformation about vaccines or abortion, and we do not allow misinformation about voting. These policies can be applied to a wide range of content, and that's by design; this content is constantly changing, often based on what's happening in the world.

In addition to removing content that is inaccurate and harms our users or community, we also remove accounts that seek to mislead people or use TikTok to deceptively sway public opinion. These activities range from inauthentic or fake account creation, to more sophisticated efforts to undermine public trust. These actors never stop evolving their tactics, and we continually seek to strengthen our policies as we detect new types of content and behaviors.

Enforcing our policies

At TikTok, a combination of technology and thousands of safety professionals work together to enforce our Community Guidelines. To do this effectively at scale, we continue to invest in technology-based flagging as well as moderation. We rely on

automated moderation when our systems have a high degree of confidence that content is violative so that we can expeditiously remove violations of our policies.

However, misinformation is different than other content issues. Context and fact-checking are critical to consistently and accurately enforcing our misinformation policies. So while we use machine learning models to help detect potential misinformation, ultimately our approach today is having our moderation team assess, confirm, and remove misinformation violations. We have specialized misinformation moderators who have enhanced training, expertise, and tools to take action on misinformation. This includes direct access to our fact-checking partners who help assess the accuracy of content.

We have more than a dozen fact-checking partners around the world that review content in over 30 languages. All of our fact-checking partners are accredited by the International Fact-Checking Network as verified signatories of the International Fact-Checking Network's code of principles. Out of an abundance of caution, while content is being fact-checked or when content can't be substantiated through fact-checking, it becomes ineligible for recommendation into For You feeds. If fact-checkers confirm content to be false, we may remove the video from our platform or make the video ineligible for recommendation into For You feeds.

To continually improve detecting and removing misinformation, we've made some key investments this year, including:

- continued investment in machine learning models and increased capacity to iterate on these models rapidly given the fast changing nature of misinformation.
- improved detection of known misleading audio and imagery to reduce manipulated content.
- a database of previously fact-checked claims to help misinformation moderators make swift and accurate decisions.
- a proactive detection program with our fact-checkers who flag new and evolving claims they're seeing across the internet. This allows us to look for these claims on our platform and remove violations. Since starting this program last quarter, we identified 33 new misinformation claims, resulting in the removal of 58,000 videos from the platform.

While violations of our integrity and authenticity policies make up less than 1% of overall video removals, these continued investments have brought gains to our proactive detection and enforcement of these policies.

Collaborating with others

We firmly believe collaborating with experts is key to tackling this challenge. We regularly engage with our Content Advisory Council, researchers, civil society organizations, and media literacy experts. Not only does this collaboration help strengthen our policies and overall knowledge of trends and issues, it also enables us to elevate authoritative voices in our app. For instance, we created a digital literacy hub in our app with content aimed at equipping people with skills to evaluate content they consume, featuring content from experts at MediaWise and the National Association of Media Literacy Education. To promote election integrity, we create Elections Centers to provide access to authoritative information about voting and local elections from organizations like the National Association of Secretaries of State and Ballotpedia. To support public health, people can access information about COVID-19 and monkeypox from the World Health Organization through hashtag PSAs and labels on videos.

We will continue to iterate on how we tackle this challenge, and we know our work will never be finished. To learn more about our efforts to safeguard our platform, visit our Transparency Center.

PX0406

Meta

Meta for Media

Tools

Blog

Resources

Log In

Where We Have Fact Checking

In certain countries, we work with independent third-party fact-checkers to review and rate the accuracy of stories.

Learn which countries our partners are working in by exploring the map or referring to the listings below.

Meta

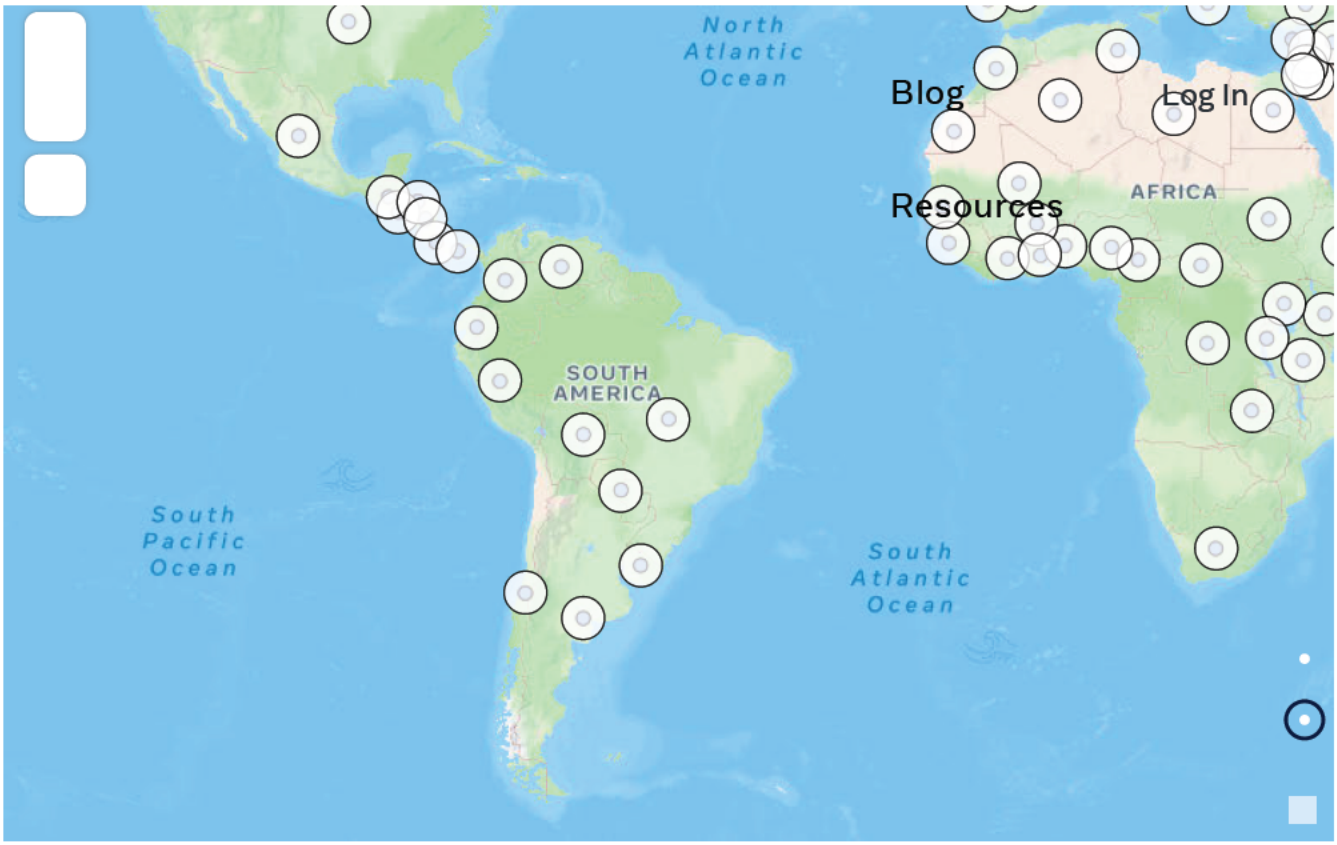
Meta for Media

Tools

Blog

Log In

Resources



A list of our independent fact-checking partners, by country

+ Albania (Albanian)

+ Algeria (Arabic)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Australia (English)

+ Austria (German, Dutch, French)

+ Azerbaijan (Turkish, Azeri)

+ Bahrain (Arabic)

+ Bangladesh (Bengali, English)

+ Belarus (Russian)

+ Belgium (Dutch, French, German)

+ Benin (French, Yoruba)

+ Bolivia (Spanish)

Meta

Meta for Media

Tools

Blog

Log In

Resources

+ Brazil (Portuguese)

+ Bulgaria (Bulgarian)

+ Burkina Faso (French)

+ Burundi (English, French)

+ Cameroon (English, French)

+ Canada (English, French)

+ Central African Republic (French)

+ Chile (Spanish)

+ Colombia (Spanish)

Meta

Meta for Media

Tools

Blog

Log In

Resources

+ Cote d'Ivoire (English, Swahili, Amharic, French, Oromo)

+ Croatia (Croatian)

+ Cyprus (Greek)

+ Czech Republic (Czech)

+ Denmark (Danish)

+ Democratic Republic of Congo (French)

+ Ecuador (Spanish)

+ Egypt (Arabic)

+ El Salvador (Spanish)

Meta

Meta for Media

Tools

Blog

Log In

Resources

+ Ethiopia (Amharic, Oromo, Tigrinya)

+ Finland (Finnish)

+ France (French, English)

+ Georgia (Georgian, Armenian, Azerbaijani, Russian)

+ Germany (German, Dutch, French)

+ Ghana (English)

+ Greece (Greek)

+ Guatemala (Spanish)

+ Guinea-Conakry (French)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Hong Kong (English, Chinese)

+ Hungary (Hungarian)

+ India (English, Hindi, Bengali, Telegu, Kannada, Malayalam, Urdu, Punjabi, Assamese, Manipuri/Meitei, Marathi, Gujarati, Tamil, Bhojpuri, Oriya, Kashmiri, Nepali)

+ Indonesia (Indonesian)

+ Iraq (Arabic)

+ Ireland (English)

+ Israel (English, Hebrew, Arabic)

+ Italy (Italian)

Meta

Meta for Media

Tools

Blog

Log In

Resources

+ Kazakhstan (Kazakh, Russian, Mongolian)

+ Kenya (English, Swahili)

+ Kosovo (Albanian, Serbian)

+ Kuwait (Arabic)

+ Latvia (Lithuanian, Russian, English, Latvian)

+ Lebanon (Arabic)

+ Libya (Arabic)

+ Lithuania (Lithuanian, Russian, English)

+ Luxembourg (German, Dutch, French)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Mali (French)

+ Mexico (Spanish)

+ Moldova (Romanian)

+ Mongolia (Mongolian, Kazakh, Russian)

+ Montenegro (Bosnian, Croatian, Serbian, Montenegrian)

+ Morocco (Arabic)

+ The Netherlands (German, Dutch, French)

+ New Zealand (English)

+ Nicaragua (Spanish)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Nigeria (English, Hausa, Yoruba, Igbo)

+ North Macedonia (Macedonian, Albanian)

+ Northern Ireland (English)

+ Norway (Norwegian)

+ Oman (Arabic)

+ Pacific Islands (English)

+ Palestine (Arabic)

+ Pakistan (English, Urdu)

+ Panama (Spanish)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Peru (Spanish)

+ Philippines (English, Tagalog)

+ Poland (Polish)

+ Portugal (Portuguese)

+ Qatar (Arabic)

+ Romania (Romanian)

+ Russia (Russian)

+ Saudi Arabia (Arabic)

+ Senegal (French, Wolof)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Singapore (English)

+ Slovakia (Slovak)

+ Slovenia (Slovene)

+ Somalia (English, French)

+ South Africa (Afrikaans, Zulu, English, Sotho, Northern Sotho, Southern Ndebele, Setswana)

+ South Korea (Korean, English)

+ Spain (Spanish, Catalan)

+ Sri Lanka (English, Tamil, Sinhalese)

+ Sudan (Arabic)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ Switzerland (German, Dutch, French)

+ Syrian Arab Republic (Arabic)

+ Taiwan (Chinese)

+ Tanzania (English, Swahili, Amharic, French, Oromo)

+ Thailand (Thai)

+ Tunisia (Arabic)

+ Turkey (Turkish, Azeri)

+ United Arab Emirates (Arabic)

+ Uganda (English, Swahili, French, Oromo)

Meta

Meta for Media

Tools

Blog

Resources

Log In

+ United Kingdom (English)

× United States (English, Spanish)

AFP - Hub
Check Your Fact
Factcheck.org
Lead Stories
PolitiFact
Science Feedback
Reuters Fact Check
TelevisaUnivision
The Dispatch
USA TODAY

+ Uruguay (Spanish)

+ Venezuela (Spanish)

+ Western Sahara (Arabic)

+ Yemen (Arabic)

Meta

Meta for Media

Tools

Blog

Log In

Resources

Meta

Sites

Meta Audience Network

Meta for Business

Meta for Creators

Meta Elevate

Facebook Gaming

Resources

Social Editorial Calendar

Media Academy

Media Support Portal

Help Center

Blog

Follow Us

© 2024 Meta

English (US)

Privacy

Terms

Cookies

PX0407

Combating Harmful Misinformation

TikTok is where over a billion people create and share content on topics that matter to them. We work hard to maintain a safe, authentic space where people can discover content that is original and engage with people who are authentic. This post explains how we encourage authentic expression on TikTok by **protecting** our community from harmful misinformation, **partnering** with experts (including our Global Fact-checking Program) to respond to evolving misinformation trends, and **empowering** our community with authoritative information.

Protecting people from harmful content

In a global community, it is natural for people to have different opinions, but when it comes to topics that impact peoples' safety, we seek to operate on a shared set of facts and reality. Our Community Guidelines prohibit inaccurate, misleading, or false content that may cause significant harm to individuals or society—regardless of the poster's intent. This includes not only physical harm, but also societal harm, like the undermining of trust in elections or public health initiatives.

Our misinformation policies

We have robust policies around specific types of misinformation like medical, climate change, and election misinformation, as well as misleading AI-generated content, conspiracy theories, and public safety issues like natural disasters. In keeping with our Community Principles (which embody our human rights commitment to protecting expression while preventing harm), our policies outline a range of enforcement outcomes, which are proportionate to how much harm the content can cause. This includes removing content or reducing its reach by making it ineligible for For You feeds. We continuously consult with experts and our global Safety Advisory Councils to ensure our approach is updated, balanced, and respectful of local nuance.

Transparency around content moderation is critical. When we take action on someone's content for violating our misinformation policies, we let them know through an Inbox notification as well as the "Account Status" page in our Safety Center, where they can also file an appeal.

"Disinformation" vs. misinformation

Our misinformation policies apply to content regardless of the poster's intent, as the content's harm is the same either way. Hence, they cover both "disinformation" (which is intentionally shared to mislead) and harmful misinformation that may not have been shared with the goal of deceiving people.

We also address disinformation by removing accounts that repeatedly post misinformation that violates our policies, and have expert teams who continuously monitor for disinformation campaigns, inauthentic behavior, and influence operations. Learn more about their work here.

Like others in our industry, we do not prohibit people from sharing personal experiences, simply inaccurate myths, or misinformation that could cause reputational or commercial harm, in order to balance creative expression with preventing harm.

Detecting violative content

We detect misinformation through automated technology, user reports, and proactive intelligence briefings from experts and our fact-checking partners. When high-risk events like elections, public safety emergencies or natural disasters are unfolding, we launch additional detection efforts locally, often collaborating with experts or authorities and establishing reporting channels for authoritative local partners to better stay ahead of evolving misinformation. When new misinformation claims are discovered, we proactively check our platform for any similar content.

Unverified content

Information evolves rapidly and sometimes it's not clear whether a claim is true or false. When we can't verify whether content is true or not, we may label it as "unverified" and reduce its spread by making it ineligible for For You feeds. This includes content about unfolding events where details are still emerging. We assess the accuracy of content by partnering with independent, International Fact-Checking Network-accredited fact checking organizations through our Global Fact-Checking Program.

Integrity & Authenticity moderators

While some misinformation can be enforced through technology alone—for example, repetitions of previously debunked content—misinformation

evolves quickly and is highly nuanced. That's why we have dedicated teams and processes for Integrity and Authenticity policies like misinformation, including enhanced training, tooling and expertise. When potential misinformation is detected, our Integrity & Authenticity moderators must determine whether it's accurate, false, or unverifiable in order to apply our policies. To do so, they consult a global database of previously fact-checked claims and then route any new, evolving, or borderline claims to our Global Fact-Checking Program for independent evaluation. Learn more about the program below.

Partnering with experts

TikTok partners with experts across the world to support consistent and accurate moderation, understand local context, and empower our community with authoritative information.

TikTok's Global Fact-checking Program

When potential misinformation is detected, our Integrity and Authenticity moderators must determine whether it's accurate, false, or unverifiable in order to apply our policies. To do so, they consult a global database of previously fact-checked claims and then route any new, evolving, or borderline claims to our Global Fact-Checking Program for independent evaluation.

Through our Global Fact-Checking Program, we partner closely with 18 IFCN-accredited fact-checking organizations who assess the accuracy of content on TikTok in over 50 languages and 100 countries across the world. These partners do not moderate content directly on TikTok, but assess whether a claim is true, false, or unsubstantiated so that our

moderators can take the right action based on our Community Guidelines. They also share intelligence that helps us detect harmful misinformation and anticipate misinformation trends.

Here's how our Global Fact-Checking Program works:

1. We detect potential harmful misinformation. In order to apply our Community Guidelines, we must confirm that it is indeed misinformation.
2. To do so, moderators consult our database of previously fact-checked claims. If the content in question represents a new or evolving claim, they route it to our fact-checking partners for assessment.
    - While the content is being fact-checked, we may make it ineligible for the For You feed out of an abundance of caution.
3. Once fact-checkers have assessed the content, our moderators apply our Community Guidelines accordingly.
    - **If content is determined to be accurate**, it stays on the platform and is eligible for For You feeds (as long as it doesn't violate any other Community Guidelines).
    - **If content is confirmed as harmful misinformation that violates our Community Guidelines**, our moderators apply our policies and either remove the video or restrict its reach.
    - If **content can't be verified after fact-checking,** we may label it as unverified and make it ineligible for For You feeds to reduce its reach.

When choosing our fact-checking partners, we prioritize organizations that are IFCN-certified, employ locally-based journalists, and specialize in assessing misinformation on social media and in short-form videos. Many

of our fact-checking partners cover several countries, which helps us scale the program to support global moderation. We currently work with: Agence France-Presse (AFP), Animal Político, Australian Associated Press (AAP), Code for Africa, dpa Deutsche Presse-Agentur, Demagog, Estadão Verifica, Facta, Lead Stories, Logically Facts, Newschecker, Newtral, Poligrafo, PolitiFact, Reuters, Science Feedback, and Teyit.

Global Fact-Checking coverage:

Leaning into local nuance

Online misinformation is a shared global challenge, but it manifests differently across countries, languages, and communities. We're continuously partnering with both creators and local experts like advocacy organizations, public health authorities, electoral commissions, and local fact-checking experts (in addition to our Global Fact-Checking Program) to reduce misinformation, elevate authoritative information, and make media literacy skills accessible to all of our global communities.

Empowering communities with reliable information

In addition to taking action on the content itself, we continuously work to deter misinformation proactively by empowering our community with media literacy resources that help them recognize misinformation, assess content critically, and file reports about violative content. These include:

- For some content that's unverified and potentially harmful, we prompt viewers to reconsider before sharing it.
- In addition to our unverified content labels, we provide a labeling tool for AI-generated content and blue "verified" checks to confirm the authenticity of notable accounts.
- For topics that can be vulnerable to misinformation—like health, elections, or unfolding crises—we direct searches towards authoritative information and add informational banners to relevant hashtag pages.
- We add informational banners to LIVE content.
- We label state-affiliated media to help viewers better understand sources behind content.

AI-generated content

Artificial Intelligence (AI) enables incredible creative opportunities, but can potentially confuse or mislead viewers if they're not aware content was generated or edited with AI. As more creators take advantage of AI to enhance their creativity, we want to support transparent and responsible content creation practices by investing in media literacy initiatives that empower creativity while giving people context about content they're viewing.

Our policies on AI-generated content

Our Synthetic and Manipulated Media policy requires creators to label AI-generated content that shows realistic scenes, and we were the first platform to build a tool for creators that helps them do this easily. We also prohibit AI-generated content that contains the likeness of any real private

figure, including anyone under 18, as well as synthetic media of public figures if the content is used for endorsements or violates any other policy. To increase clarity about AI-powered TikTok products, all TikTok effects that are significantly edited with AI must include “AI” in their name and corresponding effects label.

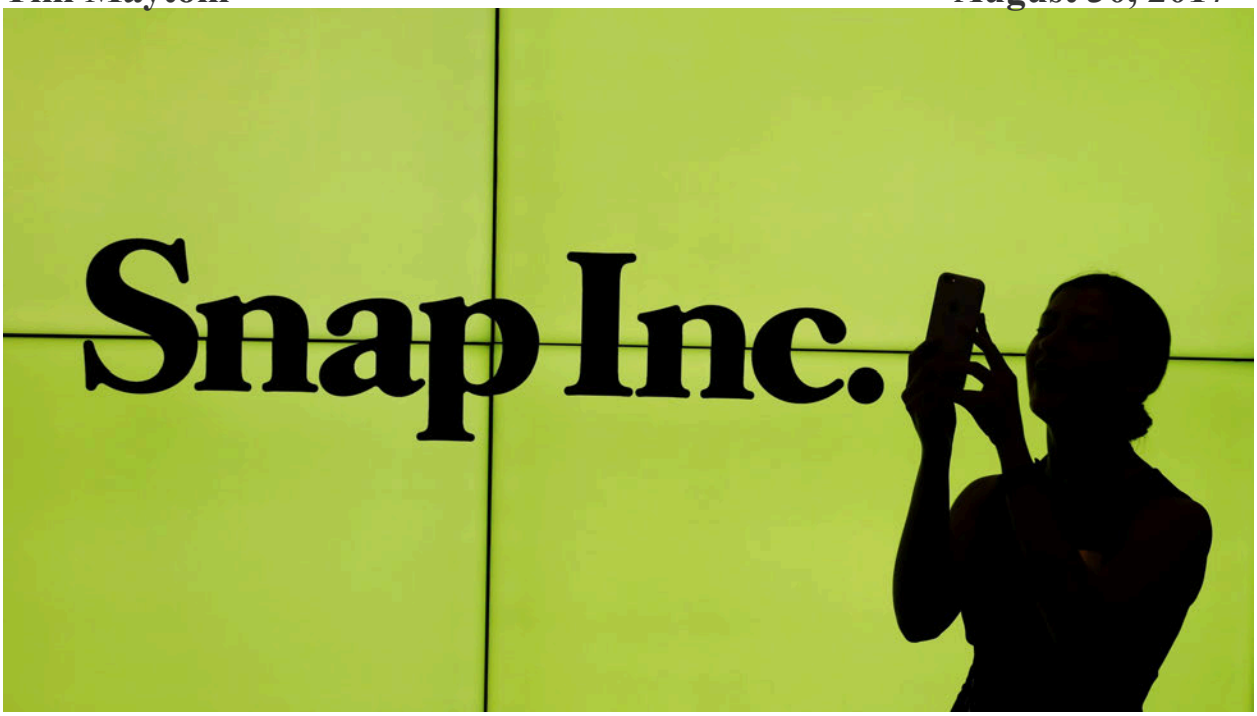
Evolving detection to keep pace with AI

As AI-generated content and technology evolves, we continue to evolve our methods of detection. This includes launching and iterating new detection methods for AI-generated content as well as assessing options for partnerships to determine the origin of content. We also continue to partner closely with experts. In February 2023, TikTok was a launch partner to the Partnership on AI’s Framework for Responsible Practices for Synthetic Media, a new code of industry best practices that promotes transparency and responsible innovation around AI-generated content.

PX0408

Snapchats in-house team of journalists are fact-checking content

Tim Maytom **August 30, 2017**



As more and more citizen reportage happens over social media, the opportunity for fake news to slip through and misrepresent events or even fabricate entire stories is growing. As many digital firms turn to tech-based solutions, Snap has instead opted for an in-houses team of journalists and fact-checkers to review uploaded content that concerns major news stories.

Speaking to BBC Radio 4s *World at One* programme, Nick Bell, vice president of content at Snap, discussed how the platform not only provided mainstream media outlets with a new channel to reach young consumers, but also served to gather user-generated content from events like the recent protests in Charlottesville, Virginia.

"We have a news team – we have journalists who work at Snapchat, who are looking at content that comes in and are evaluationg it, are determining whether it is accurate, whether it is relevant and how we can add additional context.

"So if you look at some of the events in Charlottesville, for example, in recent weeks, we actually received Snaps from members of our community of the driver being arrested. Before we published these Snaps, we actually verified with the police to make sure that it was accurate, that the snaps that we were posting to our 173m daily active users were of what we thought they were. And we added a layer of context adove that as well to describe what had happened, who this person was."

The approach is considerably different to rival social network Facebook, which disbanded its editorial curation team last year in favour of an algorigthm-driven system after accusations of political bias. The social network now relies on machine learning, as well as external services, to fight back against the tide of fake news.

While Snapchats audience is significantly smaller than Facebooks, it is growing faster, and recent figures from eMarketer suggest that young people in particular are using Snapchat and Instagram over Facebook.

"We think that actually putting journalists into the fold is very, very important, and one of the things that we think were really lucky to have is 173m people every single day who are sharing what they see in the world around them, and we can package these up into really powerful stories," said Bell.

"The shows and the content that we program on Snapchat is highly curated. We love the concept of actually taking people on a journey, having very credible perspectives who guide you through the topics of the day."