**United States Senate Committee on Homeland Security & Governmental Affairs**
**"Social Media's Impact on Homeland Security: Part II" Hearing**

**Testimony of Jay Sullivan**
**General Manager of Bluebird (Consumer Products) &**
**Interim General Manager of Goldbird (Revenue Products)**
**Twitter, Inc.**

**September 14, 2022**

Chairman Peters, Ranking Member Portman, and Members of the Committee:

Thank you for the opportunity to speak to you today on Social Media's Impact on Homeland Security. My name is Jay Sullivan. I joined Twitter last November as a Vice President on the Consumer Product team. Five months ago, I was promoted to the General Manager of Bluebird, Twitter's Consumer Product team that is responsible for the main features people use on Twitter's mobile app and website. I am also currently the interim General Manager of Goldbird, Twitter's Revenue Product team.

Twitter's purpose as a company is to serve the public conversation. The open nature of our service gives a voice to a world of diverse people, perspectives, ideas, and information. We foster free and global conversations that allow all people to consume, create, distribute, and discover information about the topics and events they care about most. At Twitter, we operate with the belief that together, we are and will continue to be a force for good in the world.

For example, in the past year we have seen people come to Twitter to get on-the-ground information about the conflict in Ukraine, including ways people can help those in need. Countless individuals have used our service to access potentially life-saving information during natural disasters, and to exchange ideas about diverse topics ranging from news, to culture, to sports.

It is our fundamental belief that public conversation should be healthy and safe. Providing our valuable service comes with challenges. As with any tool, some people and organizations will try to abuse it for their own gain or to harm others. We take our responsibility to address these issues seriously. My role at Twitter is to lead its product vision, strategy, and execution. Today, I want to focus my testimony on how keeping our service healthy and safe is an integral part of our product strategy and is essential for how we grow Twitter.

I look forward to sharing with the Committee some of the important work we are doing through product design and interventions, policies, and external engagements to make sure that Twitter is enjoyed by everyone in safe and healthy ways, and in ways that further the values of freedom of speech and expression.

## Incentivizing Health & Safety

I want to make clear at the outset that Twitter as a company is incentivized to keep our platform healthy and safe. Indeed, my top two objectives as the General Manager of Bluebird are to develop products to grow the number of users on Twitter and to prioritize health and safety. These two priorities go hand-in-hand because if people aren't protected from hate, abuse, and harassment, they will leave the service. Toxic behavior therefore impacts not only the health and safety of Twitter, but also harms long-term user growth. We therefore build health and safety into the design of new features, but if we are not satisfied, we will pause, delay, or stop a product rollout because of health and safety concerns.

A healthy and safe public conversation is also essential for our advertisers, who want to ensure that their brand, products, services, and activity are not depicted alongside harassment, vitriol, extremism, or false and misleading information. If they do not have these assurances, they will withdraw their ads from our service. Focusing on a healthy and safe public conversation also allows Twitter to further our core value of defending and respecting the rights of people using our service while promoting the core tenets of free expression.

In sum, mitigating risks and prioritizing a healthy and safe Twitter is good for our users, our business, and society. And, it allows us to better achieve our growth and financial goals..

## Product Design To Promote Healthy and Safe Public Conversation

Since health is an integral part of our product strategy and development process, Twitter collectively works to prioritize health and safety every step of the way, from ideation to launch of a product's lifecycle. At the outset, during the ideation and design phase before we begin development,, we assess the impact on health and safety. This process includes a comprehensive assessment of potential risks and unintended consequences. It also includes developing mitigation strategies, which are integrated into product development and planning. To do this work effectively, the product team works collaboratively internally and includes individuals with a range of expertise, ranging from research to human rights.

Before any major product or policy launch, a cross-functional group of people will work together to consider potential risks, unintended consequences, responses from bad actors, and risk of abuse. This work includes an analysis of the potential risk posed by product features to Twitter users, the platform, and society. Our Trust & Safety team includes subject matter experts on different issues, as well as experts focused on considering the consequences of product and policy decisions and how best to remediate them. Furthermore, our research team uses internal and external studies to inform our work, and we also have a number of employees publishing academic papers on their work to share insights.

We also build in outside perspectives in a range of ways — as part of a formal public feedback process, direct engagement with experts, academics and civil society groups, or through our research work. These perspectives help us understand risks, mitigations, and trade-offs that inform our wider product and policy strategy. We act on these risks and build mitigations as we develop products.

We know we won't get everything right the first time and that people might react differently to how we expect when a new feature or policy is launched. So we use experiments to test new features, sometimes based on geography, other times with a random sample of people around the world. We pause, delay, or cancel initiatives if the risks can't be mitigated.

At Twitter, we also have intentionally prioritized openness and real, meaningful transparency. Transparency is central to how we build and ship products, as well as how we work to improve experiences on Twitter. When we develop a new feature, as much as possible, we do so in the open, incorporating feedback from the people who use Twitter, and ensuring we create a safe, accessible end-product for everyone.

Transparency also means accountability — owning our mistakes and correcting them. When we get something wrong, we communicate transparently about it and hold ourselves accountable for fixing it. Our commitment to transparency is likewise embodied in other key ways, like our open application programming interfaces — also known as APIs — including our free academic research track access, our ongoing industry-leading disclosures of state-linked information operations, and the information regularly shared in the Twitter Transparency Center.

In addition to building health into product design, we also use product interventions to promote health and safety. We have been proactively developing a new set of products and features that give users more control over their experience and help them feel safe and we are seeing promising data and outcomes. Here are a few that are emblematic of this focus:

- *Labels on Tweets:* This allows us to label Tweets that are misleading with clear warnings, accessible context, and de-amplify and limit engagements on certain Tweets through Likes, Retweets, and Replies.

- *Prompts before certain actions like Retweets and Replies are taken:* We've found that simple prompts that encourage people to read articles — past the headline alone — or consider a potentially abusive response before sharing have a demonstrated impact. These are speed bumps that essentially slow down content creation or viral sharing, with the intended effect of encouraging people on Twitter to *consider* what they're reading or saying before sharing it.

- *Conversation Controls:* We allow people on the service to control who can respond to their Tweets. We also allow them to hide unwanted replies to their Tweets or unmention themselves from a conversation to help people have more control over their experience.

- *Birdwatch*: Birdwatch allows people to identify information in Tweets they believe is misleading and write notes that provide informative context. We believe this approach has the potential to respond quickly when misleading information spreads, adding context that people trust and find valuable. Eventually, we aim to make notes visible directly on Tweets for the global Twitter audience, when there is consensus from a broad and diverse set of contributors.

- *Disabling Algorithmic Ranking*: The Sparkle button — which has been a feature of our service since 2018 — allows people on Twitter to view Tweets in reverse chronological order, rather than in an order suggested by our technology. This is a simple tool, and gives people control.

As we develop and expand our product roadmap moving forward we will continue to build on these and introduce new capabilities to keep our platform and customers safe. These capabilities include expanding our systems and processes to ensure that we are de-amplifying objectionable content on the platform, removing content that violates our policies, and making it easier for customers to report problematic content to us.

## Policies Designed to Mitigate Harm and Promote Safety

While our Trust & Safety team is responsible for developing the policies and governance frameworks that prevent and mitigate harm to the people who use Twitter, I want to briefly touch upon our Twitter Rules — the policies we have in place to make sure people can participate in the public conversation freely and safely. These policies make clear that violence, hateful conduct, harassment, and other types of threatening behavior are not permitted on Twitter.

Our policies are built around the promotion and protection of fundamental human rights, including freedom of expression, safety, and privacy. These rights, among others, are enshrined in the Universal Declaration of Human Rights, which is an international document adopted by the U.N. and numerous countries around the world.

We believe deeply in and advocate for freedom of expression and open dialogue. We know that people do not always agree. The ability to dissent, to share information and opinions freely, even when unpopular, provocative, or questioned, is a value that makes up the foundation of free expression, but that means little as an underlying philosophy if voices are silenced because people are afraid to speak up due to threats to their physical safety or privacy. That is why we have policies that make clear that we will not allow for the promotion of violence, disinformation, or hateful conduct on Twitter as they undermine our ability to serve the public conversation, our customer experience, our business, and our ability to promote the open internet.

### *Platform Integrity & Authenticity Policies*

Our platform integrity and authenticity policies promote the health of the public conversation by addressing, among other things, efforts to spread misinformation relating to civic integrity, moments of crisis, COVID, and synthetic and manipulated media. We are constantly reviewing and evaluating misinformation efforts and focusing on those that are most harmful. For example, we added our crisis misinformation policy in May 2022 as we determined that in times of crisis — such as situations of armed conflict, public health emergencies, and large-scale natural disasters — false and misleading information has a special capacity to bring harm to vulnerable populations and shape crisis dynamics.

Our coordinated harmful activity policy addresses those situations where we find groups, movements, or campaigns that are engaged in coordinated activity resulting in harm on and off of Twitter and take enforcement action on any accounts that we identify as associated with those entities. In order to take action under this policy, we evaluate these groups, movements, or campaigns against an analytical framework, with specific on-Twitter consequences if we determine that they are harmful. You can read more about this approach here.

*Safety Policies*

Our safety policies are built to prohibit abuse, harassment, violence, and criminal actions on Twitter. Among the policies included in this category are: non-consensual nudity, suicide and self-harm, perpetrators of violent attacks, private information, hateful conduct, sensitive media, abusive behavior, violent organizations, violent threats, glorification of violence, abusive profile information, child sexual exploitation, and illegal or certain regulated goods or services. You can read more about these policies here.

Our violent threats, wishes of harm, and glorification of violence policies prohibit content on Twitter that promotes, incites, or threatens violence off of the platform. All forms of incitement of violence — whether veiled, coded, or opaque — fall squarely under this prohibition.

*Brand Safety Policy*

Our Brand Safety policy, which is led by our Customers team, builds upon the foundation laid by the Twitter Rules to promote a safe advertising experience for all users and brands. In addition to our Brand Safety efforts, which help inform ad placement on Twitter, we also have Advertising Policies that determine permissible content in ads and conduct of advertisers on Twitter. You can learn more about our Ads Policies here.

## External Engagements

We know at Twitter that our efforts to create healthy conversation require engagement with industry, academia, the public, governments, and civil society, among others, to be successful and to address the most serious online threats and develop products and policies that further our efforts. I want to highlight a few of our external engagements today.

*Twitter Trust and Safety Council*

The Twitter Trust and Safety Council is a group of independent expert organizations from around the world. Together, they advocate for safety and advise us as we develop our products, programs, and rules. At the end of 2019, we expanded the Council to include even more global experts and diverse perspectives. The Council is made up of several advisory groups, each dedicated to issues critical to the health of the public conversation.

This year we've engaged with the Trust and Safety Council on 6 projects, following the 13 projects we engaged with the Council on in 2021, early in the development process. We distilled and put to use their feedback on ways we can offer a better and safer experience for people using

Twitter. Their feedback directly informed our approach on several products. You can read more about the Twitter Trust and Safety Council here.

*Global Internet Forum to Counter Terrorism ("GIFCT")*

Twitter co-founded GIFCT with YouTube, Microsoft, and Facebook in 2017. GIFCT helps technology companies, government, civil society, and academia share information to counter terrorist and violent extremist activity online. GIFCT evolved with the Christchurch Call to Action, an initiative that governments, tech platforms, and civil society organizations committed to after the March 2019 mosque shootings in Christchurch, New Zealand and viral spread of the perpetrator's live-streamed video of the attack. Among GIFCT's work:

- A real-time content incident protocol (CIP) that allows us to respond to a violent act quickly to ensure that we share valuable information across industry to limit the spread of terrorist and violent extremist content.
- A shared, safe and secure industry database of "perceptual hashes" of known images and videos — produced by terrorist entities on the United Nations designated terrorist groups lists.
- Establishing the Global Network on Extremism and Technology (GNET), an independent, industry-funded initiative for better understanding, and counteracting, terrorist use of technology.

Since the attack in Christchurch, GIFCT members have shared alerts relating to hundreds of incidents around the world and activated its content incident protocol three times in response to violent attacks in Halle, Germany (2019), Glendale, Arizona (2020) and Buffalo, New York (2022). Supported by table-top exercises and post-incident reviews, we continue to strive to limit the spread of perpetrator-produced content, whether through video, audio or manifestos. You can read more about GIFCT here.

*Digital Trust and Safety Partnership*

The Digital Trust and Safety Partnership (DTSP) is an initiative focused on promoting a safer and more trustworthy internet. Twitter joined as an inaugural member of the group because we supported its efforts to develop, use and promote industry best practices, reviewed through internal and independent third-party assessments, to ensure consumer trust and safety when using digital services. We have committed to DTSP's five fundamental areas of best practices, including: product development, governance, enforcement, improvement, and transparency. You can read more about the commitments and self-assessments of the DTSP here.

*Global Alliance for Responsible Media (GARM)*

GARM is a cross-industry initiative established by the World Federation of Advertisers to address the challenge of harmful content on digital media platforms and its monetization via advertising. Through our engagement with the GARM, for example, we have been able to contribute to the creation of the industry-standard Brand Safety Floor and Suitability Framework.

These are but a few examples of our external engagements where we are learning, discussing, and taking action that not only promote the public conversation on Twitter but working across industry to create sharing-mechanisms and standards for keeping content safe for people across the digital ecosystem.

## Conclusion

We have invested, and will continue to invest heavily, in building the technologies, policies, and procedures necessary to offer informative and safe experiences for the millions of people on Twitter.

Our task is not an easy one and what I described today in terms of our products, policies, and external engagements will change as new challenges, risks, and threats develop. In order to mitigate the harms this Committee is examining, we must constantly change and evolve. We are committed to doing what is necessary to continuously foster a healthy service while upholding the tenets of free expression, which is, in our view, the best way for Twitter to help protect democracy in the United States and abroad.

Thank you, and I look forward to answering your questions.