Attachment—Additional Questions for the Record

Subcommittee on Communications and Technology and
Subcommittee on Consumer Protection and Commerce
Joint Hearing on
"Fostering a Healthier Internet to Protect Consumers"
October 16, 2019

Ms. Gretchen S. Peters, Executive Director, Alliance to Counter Crime Online

The Honorable Kathy Castor (D-FL)

1. **On June 19, 2019, The Verge published an investigation into one of Facebook's content moderation sites in Tampa, FL, which is operated by the firm Cognizant. The article details allegations of appalling working conditions including sexual harassment, verbal and physical fights, theft, and general filthiness in addition to adverse mental health effects associated with the nature of their work.**

    a. **Operationally, how should tech platforms moderate their content? What role should human content moderators play? What role should technology play?**

Response: Platforms should ban all serious crime activity, and should implement systems to enforce these policies, instead of just stating them. The Alliance to Counter Crime Online believes that Congress should modify existing laws to make tech firms strictly liable for hosting specific criminal and terror content, after the firms have observed the content and refuse or fail to notify law enforcement in the country where the crime is being committed, and that unless and until this happens, many firms will take a lax approach to the issue of toxic content. Firms should implement content-control systems that combine the work of human moderators and AI technology, as well as user generated tips. The last few years have shown that. While artificial intelligence can and does identify a tremendous amount of illegal activity, this identification is futile if action is not taken to stop the actual crimes being committed. The majority of AI-identified criminal content today is simply removed from the platform, and often deleted without being communicated to law enforcement. This is literal destruction of evidence rather than moderation.

There will need to be content moderators, as well as investigative units that patrol the platforms for new trends and issues. Facebook and other firms must also be more strategic in how it invests in moderators. Currently, Facebook's content moderators do not maintain any expertise in a particular subject area, they are simply "cleaners" for the platform. Many do not even have college degrees, let alone expertise in the intricacies of organized crime and terrorism. If Facebook, and other firms, were to invest in subject matter experts (i.e. terrorism experts, wildlife experts, conflict antiquities trafficking experts) they would not only be more effective at

identifying and dismantling illicit content networks, they would also have a team of people that were trained, prepared, and capable of dealing with the heinous content they so routinely encounter.

Firms should be regulated to hand over IP addresses and other identifying data about users engaging in illicit activity on their platforms to law enforcement, much the way banks must file suspicious activity reports (SARs). When a SAR is reported to the government, there should be regulations for how long the firm must hold the data (minimum 1 year) in the event that law enforcement wants to pursue a case. Because crime networks often use multiple platforms at once to mount sophisticated scams and crimes, there should be mechanisms to facilitate and encourage information and data sharing between the large tech firms and law enforcement to coordinate.

b. What standard should a private company use to evaluate content? "Quasi constitutional", a "community standard" established by the company along the lines of other private media, other?

Response: a "community standard" established by the company along the lines of other private media should be sufficient. Many companies already have perfectly acceptable policies and standards, they just fail to enforce them.

c. Given that private companies are not governed by standards that government would be when it decides not to post content, why do content moderators have to spend so much time reviewing and in such great detail evaluating explicit, violent, or hateful content? What value is there to society and the site owner to work to ensure that such explicit, violent, or hateful content is given every opportunity to be posted?

Response: There is no value to society to this, other than news content reporting on terror attacks, etc (and news content could be tagged so it would not get flagged by AI systems). Companies claim to be careful about moderation in order to avoid violating free speech, however there is no federally mandated system in place to regulate what firms must do once illicit activity has been detected, as there is with the regards to banks and financial firms. This lack of regulation is precisely why the problem has grown to this point.

Firms can and should use AI to block the vast majority of the most violent, toxic and pornographic content, only forwarding that which appears to reflect a violation of the law (e.g. child pornography, drug sales etc.), so that it may be passed to authorities. It is shocking to learn about the extent of animal abuse and torture that occurs on the web, for example. All the major online tech platforms have "community standards" banning violent and explicit content, however many black markets hide inside groups on Facebook and other platforms, where the majority of illicit activity occurs. Moves toward greater encryption by major tech platforms threaten to make this worse, not better. Congress must find a balance between rule of law and user privacy, but the two issues do not need to be mutually exclusive.

AI systems on their own will never be enough, because people are smarter than systems, and quickly learn to skirt the settings, unless there are human moderators to evaluate specific instances.

d. This explicit, violent, or hateful content often is known to be inconsistent with the tech platform's content bylaws. Why do tech platforms, like Facebook, force content moderators to not only look at but also evaluate in great detail explicit, violent, or hateful content that is often inconsistent with the tech platform's bylaws?

Response: We ask ourselves the same question every day. Could it be that firms like Facebook realize they are earning so much money off this traffic to perceive a reason not to remove the content? We don't know. We propose Congress ask tech firms like Facebook and Google to declare what revenue they get from users who come online and engage with opioid content, for example. What we do know is that every click is a datapoint, and that data has value to Facebook. When alliance researchers identify 100 Facebook groups trafficking blood antiquities, they are looking at groups with a pool of roughly 2 million members. The question to ask Facebook is: "How much is the data of those individuals worth?"

We are of the opinion that Facebook and other tech firms have created a moderation system that, on the surface, appears to be attempting to tackle the issue of illegal and explicit content. However, our research and whistleblowers who have collaborated with us report that this very same system is seemingly intentionally designed to minimize the efficacy of those moderators.

Moderators are given a daily quota of 840 "tickets" (individual flagged items to review) for each 8-hour work day. If the moderators take a 1-hour lunch break, that leaves 7 hours of work available to reach the quota of 840 tickets. On average, the moderators have to observe, investigate, make a decision, and move on to the next ticket in 30 seconds or less in order to reach their quota. This process frequently involves reading through the user's Facebook Messenger inbox to understand the user's general activity. If the moderator decides to escalate a ticket, the moderator has to state a case to his/her supervisor as to why the ticket needs to be escalated, then make a report to a representative from Facebook who will make the final decision as to whether or not the ticket warrants escalating to law enforcement or taking further action. (Based on our sample data, 100% of the time when escalated tickets involve serious illegal activity involving users who are outside of the US and Canada, Facebook has refused to escalate the ticket.) All of this effort takes time away from the 30 second window the moderators have available to stay on track to meet their quota. Additionally, in 30 seconds or less, in all but the most obvious cases it is nearly impossible to sufficiently investigate a user's account and message traffic to understand if serious crime is occurring.

Moderators who fail to meet their daily quota are denied overtime, holidays off, and do not have their contract renewed or are fired before their contract expires. This is an incredible incentive to the moderators not to escalate any items.

Moderators that we have spoken to are graded for their efficiency and accuracy in correctly marking tickets that are either scams or not scams. There is no professional performance metric

that measures whether or not moderators remove or escalate content that involves serious criminal activity so long as no scam is occurring.

In other words, if a moderator views 840 tickets all relating to the sale of 7-year-old girls to a known human trafficker, and the moderator believes that the sales are real and are not a scam, the moderator will receive a perfect efficiency rating if the moderator does not mark or escalate any of the 840 tickets.

e. Should content moderators have more leeway to ban harmful content so they don't have to look at it over such lengthy time periods and evaluate the content in such detail?

Response: Yes. In fact, as stated earlier, AI should remove it before they have to see most of it. However, moderators are not spending lengthy time periods evaluating individual cases of graphic content, as they are systematically held to a standard of 30 seconds or less per ticket. If the user posting this damaging content is not banned, the user will continue to post and put moderators in a perpetual game of whack-a-mole. The psychological damage to the moderators comes in the form of a daily onslaught of horrible images and events, made worse by the frequency that the moderators are told that no action will be taken to help the victims in the case or to prosecute the criminals.

The main issue is reporting the serious crime to authorities. Facebook's AI system known as "Sigma" automatically removes a good amount of content that pertains to illegal sale of drugs, however moderators are told that this information does not need to be escalated to law enforcement because Sigma has already removed the content (even if the drug seller and purchaser have already made contact via Facebook.) Our alliance has debriefed multiple Facebook moderators from Cognizant and other firms who describe online "auction groups" where members can pay to be invited to watch video streams of animal torture, child abuse, bestiality, sadistic rituals and other illicit activities. Moderators told us the same people ran groups like these over and over, and even took payment sometimes using Facebook payment systems. But Facebook does not report this activity, or the identities of these users to law enforcement, unless there are children in America in the videos. Animal torture and sadistic activities involving adults may stay up – even when they are in violation of Facebook's own policies. Even if there are American children being abused in those videos, the 30 seconds or less standard of investigation ensures that a great deal of crime, regardless if the victims are American children or foreign children, will go unreported and will not get escalated. This needs to change.

f. What should industry best practices be for treating content moderators? Should Congress play a role in ensuring worker rights in this unique industry? If so, how?

Response: Basic labor law violations appear to be occurring as it relates to failure to pay hourly employees overtime pay for overtime work. These contractors should be reviewed by the Department of Labor to ensure that the employees are being afforded the basic rights of American workers. Moderator are monitored constantly, to the point that their supervisors know

precisely how many times and for how long the moderators take bathroom breaks, so the data regarding their hourly and overtime hourly work is available.

Removal of the unreasonably high 840 daily ticket quota will go a long way to ensure that moderators are able to perform their very serious job with the time and attention that each ticket deserves. We are of the opinion that the combination of witnessing traumatic events and the feeling of helplessness associated with regularly being told not to escalate serious crime to law enforcement creates a particular harm to the mental health of the moderators.

Soldiers and police officers deal with very traumatic cases on a regular basis, and while these traumatic events still take their toll, soldiers and police officers experience a tremendous amount of pride and satisfaction in their work. This satisfaction is derived from being able to help others and make a difference. If the moderators are working in an environment where they are actually able to help, then the moderators will experience a similar sense of pride and fulfillment that comes from helping to make a real difference in the world.

Creating a system that rewards moderators for finding information that leads to the arrest of human traffickers, child rapists, and drug traffickers will create a fundamental change in the experience of the moderators, changing from their current experience of trauma and futility, and shifting to an experience where their efforts are able to help the victims that they see in the tickets.

Mental health is a very serious matter, and any job that requires regular and significant exposure to incredibly traumatic content responsibly needs access to professional mental-health care. One moderator that we have spoken to specifically asked to see a psychiatrist to help cope with the graphic content the moderator was being subjected to and their supervisor denied them access to a professional mental health provider.

g. Is it common practice among tech platforms to use contractors to conduct content moderation for their sites? Why do some tech platforms use contractors to conduct content moderation for their sites? Should tech platforms do this?

Response: From what we know about this still-veiled industry, is fairly common practice for tech companies to use contractors for their sites. There are several reasons they might use contractors for what are arguably the most psychologically damaging jobs in tech: 1) It keeps the companies from being liable for the mental health of these employees, 2) it is a cost-cutting measure, hiring contractors at low rates also means the companies do not have to provide health care – an especially pressing issue when considering the emotional and psychological health of the employees.

We are of the opinion that use of contractor firms to moderate content is done specifically to create layers of liability insulation. Based on the disturbing details about how insufficiently the moderators are trained and managed, the objective of Facebook, and other firms, is clearly not to police their platform, arrest criminals, and rescue victims. Rather, the appears to be taking the minimally required action to create the illusion of addressing the issue, while maintaining an

arms-length distance between content moderation and regular firm employees. This way Facebook can blame another corporation if and when the gross inadequacy of the moderation efforts become known to the public (and Congress.)

The Honorable Lisa Blunt Rochester (D-DE)

1. **What can the federal government do to improve the capacity and ability to effectively moderate online content, including technological research?**

Response: The federal government should establish a whistleblower program focused on ensuring compliance with new regulations on illicit and toxic content. Corporate crime is often difficult to detect, and the experience with SEC, CFTC and IRS whistleblower programs is that detection and enforcement is greatly enhanced when whistleblowers are incentivized to step forward. Key elements of this program would include protections against retaliation (including the right to submit evidence to authorities anonymously and confidentiality) and financial rewards to whistleblowers when their information contributions to successful prosecutions and monetary sanctions, so that tech employees and former users of illicit content can receive rewards for coming forward with information. Fines from these programs should be allocated to a fund for law enforcement and for scholarship about illicit and toxic content.

The Honorable Richard Hudson (R-NC)

1. **We've heard today that CDA 230 was intended to be both a "sword" as well as a "shield". My colleagues before me have addressed how there are legitimate concerns that the "sword" aspect is somewhat underutilized. Over the last five years we have seen a dramatic increase in terrorist activities online. Social networking sites and other online platforms have been used by terrorist organizations to not only spread their hateful rhetoric but also as a powerful recruiting tool, including inside the United States.**

    a. **Do you believe that online platforms currently do enough to counter the spread of terrorist propaganda online?**

Response: We would argue that the "sword" aspect has been almost entirely forsaken by tech firms, who realized they could scale faster if they ignored toxic content. This has allowed terror groups and drug cartels to weaponize social media, using platforms to recruit new members, broadcast their messages on a far wider scale than they could have ever imagined before, and to target victims for threats, extortion and even murder. Facebook's own AI has even *generated* terror content, according to a whistleblower report from May 2019.[1] The company has put so little effort into wielding its "sword," that the same AI technology generated more than 100 business pages for ISIS months after Facebook was questioned on the whistleblower report in a

[1] https://www.whistleblowers.org/wp-content/uploads/2019/05/Facebook-SEC-Petition-2019.pdf

Congressional hearing.[2] The company has yet to remove these pages and has not fixed the algorithm that creates them.

Furthermore, members of our alliance have released extensive reporting about how terrorists (confirmed by human intel) are using the platform to actually raise money by selling conflict antiquities – a war crime – on Facebook.[3] These alliance members have spoken to officials at Facebook directly about the urgency of these issues, but the company has yet to change its commerce policies to address these concerns. Terrorists are not just spreading propaganda through Facebook, they are using the platform as an outlet for fundraising.

b. Do you believe CDA 230 provides an adequate framework to address this growing issue?

Response: Not at all. Times have changed. When CDA230 was written, the tech industry was in its infancy and most people connected to the Internet over a dial-up telephone. Smartphones and social media had not been invented. Today's technologies allow terrorists and criminals to spread globally at a far faster pace than ever before in history. For the health and safety of the American people, we must remove immunities for hosting criminal content.

If a child was sold into prostitution inside the privacy of a motel room, nobody would think that the owner of the motel or its employees should be prosecuted or liable for the crime of human trafficking. This is in essence what was the intent of CDA 230; that a provider of a legal service should not be liable for the illegal misuse that may be committed by a third party. This makes sense, and nobody would argue the reasonableness of this limitation of liability.

But what if that specific motel is a known and frequented location where child sex trafficking occurs? What if the motel owner and their managers all know for a fact that children are regularly being sold into prostitution inside their motel rooms? What if every single motel room has full video and audio surveillance of every second of human trafficking activity, yet 9 times out of 10 the motel owner and motel managers choose to delete the video evidence and order their employees not to talk about what they saw and not to report it to law enforcement. What if the motel owner and managers are specifically receiving compensation directly related to the prostitution of children on their premises? Does anyone think that the motel owner and managers should still not be held liable?

This second situation is literally what happens on social media platforms. Facebook and Google in particular continue to use CDA 230 to avoid action and responsibility despite their available knowledge, evidence, and capability. This is why CDA 230 must be amended.

[2] https://apnews.com/3479209d927946f7a284a71d66e431c7
[3] http://atharproject.org/wp-content/uploads/2019/06/ATHAR-FB-Report-June-2019-final.pdf