# House Committee on Energy and Commerce
# *Fostering a Healthier Internet to Protect Consumers*

Hany Farid, Ph.D.

## Testimony

### Background

Technology and the internet have had a remarkable impact on our lives and society. Many educational, entertaining, and inspiring things have emerged from the past two decades in innovation. At the same time, many horrific things have emerged: a massive proliferation of child sexual abuse material [5], the spread and radicalization of domestic and international terrorists [2], the distribution of illegal and deadly drugs [10], the proliferation of mis- and dis-information campaigns designed to sow civil unrest, incite violence, and disrupt democratic elections [1], the proliferation of dangerous, hateful, and deadly conspiracy theories [9], the routine harassment of women and under-represented groups in the form of threats of sexual violence and revenge and non-consensual pornography [3], small- to large-scale fraud [12], and spectacular failures to protect our personal and sensitive data [4].

How, in 20 short years, did we go from the promise of the internet to democratize access to knowledge and make the world more understanding and enlightened, to this litany of daily horrors? Due to a combination of naivete, ideology, willful ignorance, and a mentality of growth at all costs, the titans of tech have simply failed to install proper safeguards on their services.

### The Past

The landmark case of New York v. Ferber made it illegal to create, distribute, or possess child sexual abuse material (CSAM). The result of this ruling, along with significant law enforcement efforts, was effective, and by the mid-1990s, CSAM was, according to the National Center for Missing and Exploited Children on the way to becoming a "solved problem." By the early 2000s, however, the rise of the internet brought with it an explosion in the global distribution of CSAM. Alarmed by this growth, in 2003, Attorney General Ashcroft convened executives from the top technology firms to ask them to propose a solution to eliminate this harmful content from their networks. Between 2003 and 2008 these technology companies did nothing to address the ever-growing problem of their online services being used to distribute a staggering amount of CSAM with increasingly violent acts on increasingly younger children (as young, in some cases, as a only a few months old).

In 2008, Microsoft invited me to attend a yearly meeting of a dozen or so technology companies to provide insight into why, after five years, there was no solution to the growing and troubling spread of CSAM online. Convinced that a solution was possible, I began a collaboration with Microsoft researchers to develop technology that could quickly and reliably identify and remove CSAM from online services. Within a year we had developed and deployed such a technology –

photoDNA, a robust hashing technology[1]. PhotoDNA has, in the intervening decade, seen global adoption (it is licensed at no cost) and has proven to be effective in disrupting the global distribution of previously identified CSAM: more than 95% of the nearly 18 million reports in 2018 to NCMEC's CyberTipline, constituting over 45 million pieces of identified CSAM, were from photoDNA.

This story illustrates an important point. The issue of inaction for more than five years was never one of technological limitations, it was simply an issue of will – the major technology companies at the time simply did not want to solve the problem. This is particularly inexcusable given that we were addressing some of the most unambiguously violent, heinous, and illegal content being shared on their services. The issue was, in my opinion, one of a fear. Fear that if it could be shown that CSAM could be efficiently and effectively removed, then the technology sector would have no defense for not contending with myriad abuses on their services.

**The Present**

In the intervening decade following the development and deployment of photoDNA, the titans of tech have barely done anything to improve or expand this technology. This is particularly stunning for an industry that prides itself on bold and rapid innovation.

In the defense of the technology sector, they are contending with an unprecedented amount of data: some 500 hours of video uploaded to YouTube every minute, some one billion daily uploads to Facebook, and some 500 million tweets per day. On the other hand, these same companies have had over a decade to get their house in order and have simply failed to do so. At the same time, they have managed to profit handsomely by harnessing the scale and volume of data uploaded to their services. And, these services don't seem to have trouble dealing with unwanted material on their services when it serves their interests. They routinely and quite effectively remove copyright infringement material (because of the Digital Millennium Copyright Act, DMCA) and adult pornography (which is a violation of, for example, Facebook's and YouTube's terms of service).

During his 2018 Congressional testimony, Mr. Zuckerberg repeatedly invoked artificial intelligence (AI) as the savior for content moderation (in 5 to 10 years time). Putting aside that it is not clear what we should do in the intervening decade, this claim is almost certainly overly optimistic.

Earlier this year, for example, Mike Schroepfer, Facebook's chief technology officer, showcased Facebook's latest AI technology for discriminating images of broccoli from images of marijuana [7]. Despite all of the latest advances in AI and pattern recognition, this system is only able to perform this task with an average accuracy of 91%. This means that approximately 1 in 10 times, the system is wrong. At the scale of a billion uploads a day, this technology cannot possibly automatically moderate content. And, this discrimination task is surely much easier than the task of identifying the broad class of CSAM, extremism, or dis-information material.

By comparison, the robust image hashing technique used by photoDNA has an expected error rate of approximately 1 in 50 billion. The promise of AI is just that, a promise, and we cannot wait a decade (or more) with the hope that AI will improve by nine orders of magnitude when it might be able to contend with automatic online content moderation.

In the meantime, AI and similar technologies can be used as a triage, reducing the amount of content that will eventually have to be viewed by human moderators. This, however, still poses considerable challenges given the woeful low number of moderators and the truly horrific working conditions that moderators are forced to endure [8].

[1]Robust image hashing algorithms like photoDNA work by extracting a distinct digital signature from known harmful or illegal content and comparing these signatures against content at the point of upload. Flagged content can then be instantaneously removed and reported.

The simple fact is that the titans of tech have not invested in the infrastructure, technology, or human moderation to deal with the abuses that they know occur every day on their services. The largest point of tension is that the majority of social media is driven by advertising dollars which in turn means that they are motivated to maximize the amount of time that users spend on their services. Optimizing for the number of users and user engagement is, in many cases, at odds with effective content moderation.

**End-to-End Encryption**

Earlier this year, Mr. Zuckerberg announced that Facebook is implementing end-to-end encryption on its services, preventing anyone — including Facebook — from seeing the contents of any communications [14]. In announcing the decision, Mr. Zuckerberg conceded that it came at a cost:

> *"At the same time, there are real safety concerns to address before we can implement end-to-end encryption across all of our messaging services," he wrote. "Encryption is a powerful tool for privacy, but that includes the privacy of people doing bad things. When billions of people use a service to connect, some of them are going to misuse it for truly terrible things like child exploitation, terrorism, and extortion."*

The adoption of end-to-end encryption would significantly hamper the efficacy of programs like photoDNA. This is particularly troubling given that the majority of the millions of yearly reports to NCMEC's CyberTipline originate on Facebook's Messaging services. Blindly implementing end-to-end encryption will significantly increase the risk and harm to children around the world, not to mention the inability to contend with other illegal and dangerous activities on Facebook's services.

Many in law enforcement have made the case that a move to end-to-end encryption, without allowing access under a lawful warrant, would severely hamper law enforcement and national security efforts [13]. Programs like photoDNA, for example, would be rendered completely ineffective within an end-to-end encrypted system. In response, Attorney General Barr and his British and Australian counterparts have openly urged Mr. Zuckerberg to delay the implementation of end-to-end encryption until proper safeguards can be put in place [6], as have the 28 European Union Member States[2].

We should continue to have the debate between balancing privacy afforded by end-to-end encryption and the cost to our safety. In the meantime, recent advances in encryption and robust hashing technology mean that technologies like photoDNA – robust image hashing – can be adapted to operate within an end-to-end encryption system.

Specifically, when using certain types of encryption algorithms (so-called partially- or fully-homomorphic encryption), it is possible to perform the same type of robust image hashing on encrypted data [11]. This means that encrypted images can be analyzed to determine if they are known illicit or harmful material without the need, or even ability, to decrypt the image. For all other images, this analysis provides no information about its contents, thus preserving content privacy.

[2]The 28 EU Member States recently approved by unanimity a declaration on combating the sexual abuse of children and directly addresses this issue of end-to-end encryption writing: "Offenders make use of encryption and other anonymisation techniques to hide their identity and location. They use communication platforms hosted and administered in different countries to groom children into abuse and to extort them to obtain abusive material, as law enforcement, hampered by obfuscation techniques and different legislative regimes across different jurisdictions, especially in third countries, struggles to take forward investigations. The Council urges the industry to ensure lawful access for law enforcement and other competent authorities to digital evidence, including when encrypted or hosted on IT servers located abroad, without prohibiting or weakening encryption and in full respect of privacy and fair trial guarantees consistent with applicable law."

Alternatively, robust image hashing can be implemented at the point of transmission, as opposed to the current approach where it is implemented upon receipt. In this client-side implementation, the distinct signature is extracted prior to encryption and transmitted alongside the encrypted message. Because no identifying information can be extracted from this signature, it does not reveal any details about the encrypted image while allowing for the monitoring of known CSAM and other harmful material.

**Counter-Arguments**

The argument against better content moderation and end-to-end encryption usually fall into one of several categories.

- *Freedom of expression.* It is argued that content moderation is a violation of the freedom of expression. It is not. Online services routinely ban protected speech for a variety of reasons, and can do so under their terms of service. Facebook and YouTube, for example, do not allow (legal) adult pornography on their services and do a fairly good job of removing this content. The reason they do this is because without this rule, their services would be littered with pornography, scaring away advertisers. You cannot ban protected speech and then hide behind freedom of expression as an excuse for inaction.

- *Marketplace of ideas.* It is argued that we should allow all forms of speech and then allow users to choose from the marketplace of ideas. There is, however, no counter-speech to child sexual abuse material, bomb-making and beheading videos, threats of rape, revenge porn, or fraud. And even if there was, the marketplace of ideas only works if the marketplace is fair. It is not: the online services have their thumbs on the scale because they promote content that engages users to stay on their services longer and this content tends to be the most outrageous, salacious, and controversial.

- *Sunshine.* It is argued that "sunshine is the best disinfectant," and that the best way to counter hate-speech is with more speech. This, again, assumes a fair marketplace where ideas are given equal airtime, and that the dialogue around competing viewpoints is reasoned, thoughtful, and respectful. Perhaps this is true at the Oxford debate club, but it is certainly not the case on YouTube, Twitter, and Facebook where some of the most hateful, illegal, and dangerous content is routinely shared and celebrated. Perhaps sunshine is the best disinfectant – but for germs, not the plague.

- *Complexity.* It is argued by the technology companies that content moderation is too complex because material often falls into a gray area where it is difficult to determine its appropriateness. While it is certainly true that some material can be difficult to classify, it is also true that large amounts of material are unambiguously illegal or violations of terms of service. There is no need to be crippled by indecision when it comes to this clear-cut content.

- *Slippery slope.* It is argued that if we remove one type of material, then we will remove another, and another, and another, thus slowly eroding the global exchange of ideas. It is difficult to take this argument seriously because in the physical world we place constraints on speech without the predicted dire consequences. Why should the online world be any different when it comes to removing illegal and dangerous content?

- *Privacy.* It is argued that end-to-end encryption, without safeguards or access under a lawful warrant, is necessary to protect our privacy. Erica Portnoy, from the Electronic Frontier

Foundation (EFF), for example, argues that "*A secure messenger should provide the same amount of privacy as you have in your living room. And the D.O.J. is saying it would be worth putting a camera in every living room to catch a few child predators.*" [13] On the first part, we agree: you have certain expectations of privacy in your living room, but not absolute privacy. On the second part, we disagree: First, the DOJ is not asking to place a camera in every living room. It is asking to be allowed to view content when a lawful warrant has been issued, as it can in your living room. And lastly, is the EFF really comfortable referring to 45 million pieces of child sexual abuse material reported to NCMEC last year as "a few child predators?"

### Conclusions

We can and we must do better when it comes to contending with some of the most violent, harmful, dangerous, and hateful content online. I reject the naysayers that argue that it is too difficult or impossible, or those that say that reasonable and responsible content moderation will lead to the stifling of an open exchange of ideas.

## References


[1] S. Bradshaw and P. Howard. The global disinformation order. *Computational Propaganda Research Project*, Sep 2019.

[2] M. Fisher and A. Taub. How everyday social media users become real-world extremists. *New York Times*, Apr 2018.

[3] S. Haynes. 'A toxic place for women.' a new study reveals the scale of abuse on Twitter. *Time*, Dec 2018.

[4] V. Ho. Facebook's privacy problems: a roundup. *The Guardian*, Dec 2018.

[5] M. Keller and G. Dance. The internet is overrun with images of child sexual abuse. what went wrong? *New York Times*, Sep 2019.

[6] R. McMillan, J. Horwitz, and D. Volz. Barr presses Facebook on encryption, setting up clash over privacy. *Wall Street Journal*, Oct 2019.

[7] C. Metz and M. Isaac. Facebook's A.I. whiz now faces the task of cleaning it up. sometimes that brings him to tears. *New York Times*, May 2019.

[8] C. Newton. Bodies in seats. *The Verge*, Jun 2019.

[9] B. Resnick. Social media's conspiracy theory problem isn't going away. *Vox*, Aug 2019.

[10] D. Scott. This is how easy it is to order deadly opioids over the internet. *Vox*, Jan 2018.

[11] P. Singh and H. Farid. Robust homomorphic image hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–18, 2019.

[12] S. Tompor. Scammers are fooling millennials out of millions of dollars: Here's how. *Detroit Free Press*, Oct 2019.

[13] J. Valentino-DeVries and G. Dance. Facebook encryption eyed in fight against online child sex abuse. *New York Times*, Oct 2019.

[14] M. Zuckerberg. A privacy-focused vision for social networking, Mar 2019.