



Report of the Anti-Defamation League on Confronting Cyberhate

5th Global Forum for Combating
Anti-Semitism

May, 2015

ANTI-DEFAMATION LEAGUE

Barry Curtiss-Lusher
National Chair

Abraham H. Foxman
National Director

Kenneth Jacobson
Deputy National Director

Milton S. Schneider
President, Anti-Defamation League Foundation

CIVIL RIGHTS DIVISION

Christopher Wolf
Chair

Deborah M. Lauter
Director

Steven M. Freeman
Associate Director

Eva Vega-Olds
Assistant Director

Jonathan Vick
Assistant Director, Cyberhate Response

CENTER ON EXTREMISM

Mitch Markow
Chair

Oren Segal
Director

Lauren Steinberg
Terrorism Analyst

For additional and updated resources please see: www.adl.org

Copies of this publication are available in the Rita and Leo Greenland Library and Research Center.

©2015 Anti-Defamation League | Printed in the United States of America | All Rights Reserved



Anti-Defamation League
605 Third Avenue, New York, NY 10158-3560
www.adl.org

Table of Contents

PREFACE	4
INTRODUCTION	5
CHARTING PROGRESS	6
A NEW CHALLENGE: TERRORIST USE OF SOCIAL MEDIA	11
RECOMMENDATIONS FOR NEXT STEPS	16
APPENDICES	19
APPENDIX A: BEST PRACTICES.....	19
APPENDIX B: CYBER-SAFETY ACTION GUIDE.....	27
APPENDIX C: REVISED INDUSTRY POLICIES AND PRACTICES	30

GLOBAL FORUM FOR COMBATING ANTI-SEMITISM

REPORT OF THE ANTI-DEFAMATION LEAGUE

ON INTERNET HATE – MAY 2015

PREFACE

The [Anti-Defamation League \(ADL\)](#) has been addressing the scourge of online anti-Semitism since pre-Internet days, when dial-up bulletin boards were a prominent communications tool. As the Internet emerged for personal use in the 1990's, ADL was there to monitor, report and propose approaches to fight online hate. Today, ADL is known as one of the preeminent NGO's addressing online anti-Semitism and all forms of online hate.

Five years ago, the Inter-Parliamentary Coalition for Combating Anti-Semitism (ICCA) established a task force to study the issue of online hate and to develop policy recommendations to address this growing problem. In May 2012, the ICCA Task Force asked ADL to assume a leadership role in this area by convening a Working Group on Cyberhate. ADL responded by bringing together representatives of the Internet industry, civil society, the legal community and academia. One year later, at the last Global Forum on Anti-Semitism, the Chairs of the Task Force, Speaker of the Knesset Yuli Edelstein and ADL Civil Rights Chair Christopher Wolf submitted a [report](#) to the ICCA with recommendations.

At this year's Global Forum, ADL is pleased to offer this progress report on behalf of the Task Force and the Working Group. It is intended to update and supplement the 2013 report, focusing on some significant achievements over the past two years as well as some serious new challenges.

Since its formation, the Working Group met four times. Its members have shared their experiences and perspectives, bringing many new insights and ideas to the table. Their input and guidance have been invaluable, especially in dealing with online issues not even contemplated when the original Task Force was created, such as the explosive growth of social media and the expanding use of the Internet by terrorist and extremist groups. Today, while we still must address problematic web sites, offensive reviews on e-commerce platforms and other familiar questions, the bigger challenges come from social media and use of the Internet by terrorist and extremist groups.

This report is organized in five sections: (1) Introduction; (2) Charting Progress; (3) A New Challenge: Terrorist Use of Social Media; (4) Recommendations; and (5) Appendices. The chart in Section Two illustrates how the challenges posed by cyberhate have evolved through the years, focusing on the differences since the previous Global Forum. Section Three highlights a challenge that has emerged since the last Global Forum related to terrorist use of the Internet. Section Four looks back at the recommendations from the 2013 report, assesses

progress, and offers updated recommendations. The Appendices include the Best Practices published by the Anti-Defamation League and supported by the Anti-Cyberhate Working Group; ADL's Cyber-Safety Action Guide; and statements from major companies highlighting their commitment to address the problem of hate online and summarizing some recent changes in their policies and practices

INTRODUCTION

The Internet is the largest marketplace of ideas the world has known. It enables communications, education, entertainment and commerce on an incredible scale. The Internet has helped to empower the powerless, reunite the separated, connect the isolated and provide new lifelines for the disabled. By facilitating communication around the globe, the Internet has been a transformative tool for information-sharing, education, human interaction and social change. All of us treasure the freedom of expression that lies at its very core.

Unfortunately, while the Internet's capacity to improve the world is boundless, it also is used by some to transmit anti-Semitism and other forms of hate and prejudice, including anti-Muslim bigotry, racism, homophobia, misogyny, and xenophobia. This cyberhate, defined in the 2013 report as "the use of any electronic technology to spread bigoted, discriminatory, terrorist and extremist information," manifests itself on websites and blogs, as well as in chat rooms, social media, comment sections and gaming. In short, hate is present in many forms on the Internet, unfortunately creating a hostile environment and reducing equal access to its benefits for those targeted by hatred and intimidation.

In an ideal world, people would not choose to communicate hate. But in the real world they do, all too often. And hate expressed online can lead to real-world violence, nearby or far away. Cyberhate poses additional challenges, because everyone can be a publisher on the Internet, hateful content can spread around the globe literally in seconds, and it often goes unchallenged. So we need to find effective ways to confront online hate, to educate about its dangers, to encourage individuals and communities to speak out when they see it, and to find and create tools and means to deter it and to mitigate its negative impact. In doing so, it is also important to keep in mind the need to find the right balance, which addresses cyberhate while still respecting free expression and not inhibiting legitimate debate.

The unique challenge of hate speech online, which prompted the creation of Working Group, is unfortunately not the only challenge we face today. Since the last Global Forum, extremists and terrorists have become much more sophisticated in their use of social media. This growing threat has been particularly evident with a rise in "self-radicalization," encouraged and abetted by terrorist groups. Terrorist exploitation of the Internet is an order of magnitude different from hate speech online, and new strategies may be necessary to respond to it.

CHARTING PROGRESS

The following chart has been created to illustrate the predominant Internet environment that existed at the time of the 2013 Global Forum, and to contrast that environment to the one we face today, in the spring of 2015. In snapshot form, it shows where we are today, revealing changes in both how cyberhate manifests itself on the Internet, and how the industry has become more serious and more sophisticated in dealing with the problem. It documents progress – often incremental, but in its totality significant, impressive and important – mostly the result of industry-sponsored initiatives. Much of this progress actually reflects recommendations included in the 2013 report and discussed in more detail in the recommendations section of this report, below. At the same time, the chart also makes it clear that the problem is far from resolved.

Unfortunately, spewing anti-Semitism and hate online is much easier than finding effective ways to respond to it. We have come to understand that as long as hate exists in the real world, that hate will be reflected in the virtual world as well. What happens on the Internet is a reflection of society, and not the other way around. Consequently, as long as technology keeps evolving, and bias, racism and anti-Semitism persist, the haters will likely find ways to exploit the new services and new platforms to spew their corrosive message. We need to be just as creative, and just as determined, to counter them.

This first section of the chart highlights changes in each of the various platforms that comprise the Internet when it comes to dealing with cyberhate.

PLATFORMS	2013	2015
Websites	Limited existence and enforcement of hosting company rules	Mixed picture. Many companies with appropriate terms of service are responsive. Companies with lax terms of service are less responsive. Potential impact of proposed new Federal Communications Commission (FCC) regulations considering U.S.-based hosts as "common carriers" is still to be determined.
Comments/Reviews	Sporadic enforcement of hate speech in review and comment sections of websites. Terms of Service not always clear or easily found.	Word-sifting software coming into increasingly frequent use. Websites far more responsive to complaints about issues in reviews and comments. Anonymity, which is used to hide identity of haters, increasingly is being addressed by online services.
Monetization and E-Commerce	Many hate groups used PayPal, Amazon, GoFundMe and similar services	Most transactional and funding websites have Terms of Service prohibiting use by hate groups (as defined by the company) and responsiveness increasing
Social Media	Limited prohibitions on hate speech, defamation or abusive posts	Universal acknowledgement of hate speech as a problem. More but confusing array of standards and mechanisms in use. Some companies more responsive to complaints than others.
Blogs	Lack of meaningful Terms of Service	Google launched updated and unified Terms of Service (3/2014) affecting content rules for most user generated content services. Impact TBD.
File Drops/Cloud Storage	Limited attention to use of file drops by terrorists and hate groups	Most file-drop sites only prohibit illegal content but do not monitor on privacy grounds
Smart Devices/Apps	No ratings for apps, games or other smart device content	Google play-initiated app rating and review system 3/2015. Impact TBD.
Games	Microsoft Xbox Unit virtually alone in enforcing gaming environment rules	Many online game platforms now using filter software to monitor and limit inappropriate language used within games as well as users name/profile information

The second section of the chart highlights basic industry practices related to cyberhate and shows how they have changed between 2013 and 2015.

PRACTICES	2013	2015
Hate Speech Policies	The Terms of Service for many platforms did not address hate speech directly or use vague terminology in policies	Multiple platforms including Facebook, Google, Twitter, Amazon, Microsoft gaming, Yahoo now include specific prohibition of hate speech
User-Friendly Reporting	Complaint mechanisms or contact details were often buried or limited in functionality	Virtually every major service and platform uses post, profile and image flagging. Now standard practice to send receipt of complaint acknowledgements and provide links to further policy/process information.
Enforcement Mechanisms	In cases where hate speech was prohibited, penalties were mostly delineated	Google, Facebook, Twitter have instituted flagging for specific posts and partial content removal. Several social media platforms have implemented "stop and think before sending" messages and campaigns.
Transparency	Pervasive tendency for companies not to explain why content allowed to remain after a complaint; little explanation offered to users whose material was deleted	Most platforms offer explanations to users whose content has been deleted and provide appeals process. Complainants on Facebook and YouTube are advised if content has been removed. Public disclosure of rationales for removals is limited.
Counter-speech	Counter-speech education by only limited number of companies, and un-coordinated between companies	Counter-speech initiatives and studies appearing on almost every major platform

The third section of the chart highlights challenges that the industry as a whole confronts when dealing with cyberhate, and noteworthy developments between 2013 and 2015.

INTERNAL INDUSTRY CHALLENGES	2013	2015
Industry Realities	No effort to broadly explain the challenges created by evolving technology, unintended consequences and the volume of content	Industry platforms are sharing more data on traffic, members' complaints and responses than ever before - but still falling short in adequately illuminating the enormous and ever-growing volume of content and the challenge of addressing issues that require human evaluation and intervention
Anonymity	Anonymous participation on many platforms tolerated despite policies to the contrary	Anonymity continues to pose challenges for enforcement of Terms of Service
Industry Coordination	No coordinated industry statements or projects obvious to the public	The Anti-Cyberhate Working Group has become a major venue for the industry to coordinate anti-cyberhate activity. Major breakthroughs: publications of ADL's "Best Practices for Responding to Cyberhate" and well-received Cyber-Safety Action Guide.
Hate speech links and linked material	Platforms took no substantial responsibility for third party or linked content	Ongoing debate and discussion regarding platform as publisher and impact of link distribution
Corporate Voices	Few if any corporate voices spoke about online hate	Anti-hate speech voices in industry now led by Facebook, Microsoft, and Google with recent important statements by Twitter

The fourth section of the chart highlights ongoing external challenges that impact the industry’s ability to address cyberhate.

EXTERNAL INDUSTRY CHALLENGES	2013	2015
Cross Border	Limited coordination of cross border issues	In the borderless environment of the Internet, almost all initiatives and resolution programs remain geographically based
Government Intervention	Uncoordinated or unenforceable regulations	Increasing disconnect between online ideals and achievable targets for action compared to laws under consideration and being enacted to curb online hate
Cyber-Terror/Hacking	Hacking (website defacement) mainly performed on an opportunistic basis without consistent political motivation or targeting	Sharp increase in politically motivated hacking targeting Jewish institutions and Western interests

The fifth and final section of the chart highlights activities by non-industry stakeholders to address cyberhate.

STAKEHOLDERS	2013	2015
International Bodies	Numerous country-specific orgs- few international networks or associations	Unchanged
Academia	Limited external and stakeholder events by major institutions	Centers flourishing at major universities, including Stanford, Harvard, Brandeis, UCLA and Yale in the U.S.
Industry- ICCA Anti-Cyberhate Working Group	Anti-Cyberhate Working Group-First Steps	Anti-Cyberhate Working Group continues to promote coordination among stakeholders; probably still the best hope for productive results

A NEW CHALLENGE: TERRORIST USE OF SOCIAL MEDIA

As Internet proficiency and the use of social media grow ever-more universal, so too do the efforts of terrorist groups to exploit new technology in order to make materials that justify and sanction violence more accessible and practical. Terrorist groups are not only using Facebook, Twitter, YouTube, and various other platforms to spread their messages, but also actively to recruit adherents who live in the communities they seek to target.

While the fundamental ideological content of terrorist propaganda has remained consistent for two decades – replete with militant condemnations of perceived transgressions against Muslims worldwide, appeals for violence and anti-Semitism – terrorists groups are now able to reach, recruit and motivate extremists more quickly and effectively than ever before by adapting their messages to new technology.

In the past, plots were directed by foreign terrorist organizations or their affiliates and recruitment and planning generally required some direct, face-to-face interaction with terrorist operatives. Indoctrination came directly from extremist peers, teachers or clerics. Individuals would then advance through the radicalization process through constant interaction with like-minded sympathizers or, as the 2007 New York Police Department (NYPD) report on radicalization described, with a “spiritual sanctioner” who gave credence to those beliefs.

The Internet and Self-Radicalization

Today, individuals can find analogous social networks, inspiration and encouragement online, packaged neatly together with bomb-making instructions. This enables adherents to self-radicalize without face-to-face contact with an established terrorist group or cell. Furthermore, individual extremists, or lone wolves, are also increasingly self-radicalizing online with no physical interactions with established terrorist groups or cells – a development that can make it more difficult for law enforcement to detect plots in their earliest stages.

At least 85% of the American citizens and residents linked to terrorist activity motivated by Islamic extremism since 2013 actively used the Internet to access propaganda or otherwise facilitate their extremist activity. One hundred percent of the U.S. residents linked to Islamic extremist activity in 2015 have used the Internet for those purposes.

ISIS Recruitment Online

Since 2014, the Islamic State of Iraq and Syria (ISIS) has been particularly aggressive in pursuing multiple sophisticated online recruiting and propaganda efforts. ISIS’s far-reaching propaganda machine has not only attracted thousands of recruits, but has also helped Syria and Iraq emerge as the destinations of choice for this generation of extremists.

This activity has likely contributed to the increasing number of individuals accused of joining or aiding ISIS and other terrorist organizations. Through the first few months of 2015, more than 20

American citizens and residents have been arrested on Islamic terror related charges. Since 2014, more than 40 have been linked to terror-related activity, including more women and minors than ever before.

Globally, at least 20,000 fighters are believed to have traveled to join the conflict in Syria and Iraq, many of whom have joined ISIS. The largest number come from Tunisia – there are believed to be between 1,500 and 3,000 Tunisian fighters. That number is followed by Saudi Arabia, from which the estimate is between 1,500 and 2,500. But non-majority-Muslim countries have seen steady numbers of individuals leaving to fight as well. This includes 800-1,500 from Russia, 1,200 from France, 500-600 from Germany, 500-600 from the United Kingdom, and about 300 from China.

There have also been a surprisingly large number of minors. For example, focusing on the United States, five Americans under the age of 18 were detained while allegedly attempting to join ISIS in 2014. This included three Denver, Colorado teenagers, aged 15, 16 and 17. At least one of the girls was encouraged to travel to Syria by an individual she was communicating with online, according to reports. The 15-year-old described her radicalization in a series of Tweets. “I started to notice the people I called ‘friends’ weren’t my true friends. But the people who reminded me about my *Deen* (religious path) were my TRUE friends.” Some of the 16-year-old’s Tweets reveal the degree to which she identified with this extreme ideology “Those who identify as ‘gay’ and ‘Muslim’ at the same time deserve death,” and “Muslims handing out apologies (sic) because of 9/11 are a disgrace to the *Ummah* (global community of Muslims).”

Twitter is ISIS’ platform of choice, in part because it is able to conceal the identities of its users more effectively than on forums and other social networking sites. And while accounts are regularly shut down by Twitter, new ones can almost always be immediately established.

ISIS’ Twitter presence is worldwide, and presented in multiple languages, as is the propaganda it distributes via Twitter. The terror group regularly releases magazines in Arabic, English and French, and it has also released propaganda statements and videos in other languages, including Hebrew, Spanish, Turkish, Russian, Kurdish, and German.

Official ISIS accounts are augmented by supporters, some of whom seem to have quasi-official status. These supporters both share official propaganda and contribute to the barrage of online voices supporting terrorist ideology. Some supporters add personal details about their experiences in the group – information that adds to the authenticity of their narratives by providing concrete experiences.

In order to unify its messaging, ISIS also organizes hashtag campaigns, encouraging supporters to repeatedly Tweet various hashtags such as #CalamityWillBefallUS, which threatened attacks against the U.S.; #AllEyesOnISIS, which attempted to magnify the number of ISIS supporters on Twitter; and #FightForHim, which called for copycat attacks following the 2014 attacks on the French magazine *Charlie Hebdo*. The goal is for these terms to trend on Twitter, vastly increasing the visibility of tweets.

Similarly, ISIS uses hashtag campaigns to insert its messages into other trending topics on Twitter that have nothing to do with violent extremism. Thus, it will encourage its supporters to tweet ISIS messages with popular hashtags such as #worldcup or #Ferguson so that people searching for those hashtags will inadvertently come across pro-ISIS posts. Hashtag campaigns have been conducted in a number of languages, including English, French, Arabic and Turkish.

ISIS supporters are often active on a variety of platforms beyond Twitter, including the social networking site Facebook, the picture-sharing site Instagram, the chat services Kik and WhatsApp, the video sharing site YouTube, and the question and answer service Ask.FM. These individuals also encourage direct contact with potential recruits via encrypted messaging services such as SureSpot.

On Ask.FM, where users can post questions anonymously, known members of extremist organizations are asked questions by potential recruits. For example, the user Mujahid Miski (believed to be Mohamed Abdullahi Hassan, an Al Shabaab member from Minnesota) was asked and answered questions including, “My brother wants to be a *mujahid* (fighter) but he’s got glasses. Will that stop him from becoming one?” Many of his answers also include encouragement for readers to join terrorist groups, including ISIS. In one, for example, he wrote, “every minute and every second is wasted if you’re not out there building the Islamic Caliphate (a reference to ISIS). Go out and make *hijrah* (migration to a Muslim land) from the east and the west and join *jihad* (the fighting). Let your blood be the water for the tree of *Khilafah* (caliphate, a reference to ISIS).

Many ISIS supporters also take advantage of the websites Justpaste.it and its Arabic-language counterpart Manbar.me, which enable them to quickly publish content to unique URLs online, which can then be shared on social media. ISIS supporters have used these sites to publish links to downloadable propaganda materials, instructions for traveling to Syria and Iraq, manifestos encouraging lone wolf attacks, and more.

Because terrorist accounts are regularly removed from Twitter and Facebook, terrorist groups and supporters have occasionally attempted to move to other platforms or create new ones, albeit with limited success. In July of 2014, for example, ISIS announced that its official Internet presence was moving from Twitter to alternate social media sites Friendica and Quitter. Following exposure by the ADL, however, all ISIS presence was quickly deleted from Friendica and Quitter, and the group returned to Twitter. Since at least November 2014, ISIS supporters have successfully broadcast terrorist propaganda on the website Mixlr, a platform that enables users to broadcast live audio “to the world” and to “chat, engage and interact with your listeners in real time.”

A number of ISIS supporters maintain blogs on which they detail their extremist ideology and narratives of an idealized day-to-day life, which they hope will appeal to potential recruits. There have also been instances of ISIS supporters creating new websites to make ISIS propaganda even more accessible. In February 2015, an ISIS supporter created a website called IS-Tube that featured a searchable archive of ISIS propaganda videos, including videos depicting

beheadings. The site was hosted on a Google-owned IP-bloc, and was removed after ADL alerted Google to its presence.

Other Terrorist Groups Using the Internet

Other terrorist organizations use social media as well, and many have learned from ISIS's techniques. During the 2014 conflict between Israel and Hamas, for example, ADL documented no fewer than 17 social media profiles that could be considered official Hamas accounts. Like ISIS followers, Hamas supporters utilized hashtag campaigns to promote terror attacks against Israelis and posted videos and images to social media that both applauded and encouraged killing Israelis and Jews with hatchets and by running them over.

Advances in technology have enabled terrorist video production to rival high quality Western films. ISIS even released a feature-film length video, titled "Flames of War," that portrayed the group as part of an apocalyptic struggle of good versus evil. Other terrorist groups – including Al Qaeda, Al Shabaab (Al Qaeda in Somalia), Boko Haram, Taliban affiliates, the Caucasus Emirates and more – have also distributed propaganda videos via Twitter in recent years.

Perhaps the most infamous English-language terrorist magazine, *Inspire*, is now distributed via Twitter instead of on extremist forums. An online English-language propaganda magazine produced by Al Qaeda in the Arabian Peninsula (AQAP), *Inspire* provides articles about terrorist ideology, recruitment information, and encouragement and instructions for homegrown attacks, including the very bomb-making instructions that the Tsarnaev brothers allegedly utilized in the 2013 Boston Marathon bombing.

An article in the magazine's second issue encouraged "brothers and sisters coming from the West to consider attacking the West in its own backyard. The effect is much greater, it always embarrasses the enemy, and these types of individual attacks are nearly impossible for them to contain." Its 2014 editions contained directions for making car bombs and bombs designed to evade airport security measures, as well as instructions regarding the best places to detonate them.

Outside the sphere of social media, terrorist groups and sympathizers have also attempted to create applications promoting their organizations and propaganda on iTunes and Google Play.

Hezbollah, for example, has launched a number of applications that provide streaming access to the group's propaganda-based television station, Al Manar. Google Play and iTunes have been quick to remove them, but Hezbollah, having blamed "the Jewish Anti-Defamation League" for launching a "campaign" to remove the original application, has created the applications so users can download them directly from the Hezbollah website, without going through iTunes or Google Play. Hezbollah has also created several video games on its website with the explicit intent of indoctrinating young players.

Other applications are created by terrorist supporters. The Anwar al-Awlaki application, for example, enabled users to listen to Awlaki's sermons directly from their mobile devices. Awlaki, the creator of *Inspire* magazine, was the primary English-language spokesman for AQAP until

he was killed by a drone strike in 2011. Awlaki remains tremendously influential. Many of his lectures remain available on YouTube, and supporters regularly create Facebook and Twitter profiles dedicated to sharing his quotes. When these profiles are removed, they are quickly replaced by new ones. A significant number of domestic Islamic extremists, including the Tsarnaev brothers, have accessed his propaganda and cited him as an inspiration.

Another Challenge: Islamic Extremists Hacking Activity

Perhaps the newest frontier of online extremism comes in the form of Islamic extremist hackings. Politically motivated hackers from the Arab world have begun targeting the websites of perceived supporters of Israel, including synagogues, Jewish institutions, and individuals. These attacks are increasingly undertaken in the name of terrorist organizations, particularly ISIS. There are signs that ISIS is beginning to attempt to harness the hackers and hacker groups into supporting its own mission and expanding the hacks to target websites and government institutions in the U.S.

In March 2015, for example, hacker(s) identifying as “ISIS cyber army” claimed responsibility for hacking 51 American websites on March 24. Each of the hacked websites was defaced with the ISIS flag, a statement that the website was “Hacked by Islamic State” and an e-mail address for the ISIS cyber army, the unit believed to be behind the cyber activities of ISIS. In the past, the ISIS cyber unit claimed responsibility for involvement in a series of attacks against a number of Israeli websites.

This capability to engage in cyber-attacks may be a reflection of ISIS’s calls for support from individuals with various skills, from media experts to doctors, to join and contribute to the group and its mission of gaining strength and territory however they can.

In April 2015, as international hackers once again set their sights on Jewish and Israeli targets as part of “OpIsrael,” an annual anti-Israel cyber-attack campaign, there were strong indications that AnonGhost, an international hacker group that supports terrorist groups and frequently employs anti-Semitism as part of its cyber activity, had replaced Anonymous as the main organizer of the campaign.

Groups such as AnonGhost express unambiguous support for the Palestinian terrorist group Hamas and the Islamic State (ISIS) and have carried out cyber-attacks in their names, bringing an Islamic extremist element into an already virulently anti-Israel and anti-Semitic campaign.

AnonGhost threatened individual Israelis with violence through mobile devices, claiming to have obtained personal information on more than 200 Israelis. One threatening text the group claims to have sent to an Israeli included an image of an infamous ISIS fighter with the caption, “We are coming O Jews to kill you.” A text sent to another Israeli man included an image of his family with the threat, “I’ll stick a knife in their throats.”

While anti-Semitic themes existed in previous OpIsrael campaigns, it had been primarily billed as a response to the Israeli-Palestinian conflict. AnonGhost’s participation and tactics thus far speak to the centrality of anti-Semitism in this year’s campaign, which serves as an extension of AnonGhost’s pro-terror activism around the world.

RECOMMENDATIONS FOR NEXT STEPS

This section is divided into two parts. The first part focuses on hate speech online, and the second part offers recommendations specifically related to the emerging challenge of terrorist use of social media discussed above.

Hate Speech Online: Updating the 2013 Recommendations

According to the 2013 report:

The problem of cyberhate is pervasive and, given the difficulty of responding to it, persistent. However, it is clear from the testimony received that a number of factors make a legislative response to cyberhate both inappropriate and likely to fail:

- The core value and benefits of free speech
- The location of most hate content on U.S. servers
- The extreme difficulty of responding to cyberhate (even by willing intermediaries) including scale and definition.
- The failure of cross-border law enforcement and civil actions to produce any meaningful change in the amount and intensity of cyberhate [*i.e.* law is not an effective tool to deal with the scale of online hate].
- The ever-changing technology which makes cross-border law enforcement and civil actions significantly more difficult.

In light of this, the Task Force recommended against any new legislation on cyberhate with the exception of educational efforts, adding that countries with speech codes “should use discretion in enforcing laws against Internet hate speech so as not to trivialize the law. The law should be reserved for particularly egregious cases.” At the same time, the Task Force added that “governments should ensure that laws and policing agency policy are sufficiently robust to ensure that they can respond to those actions that move beyond words into real world criminal behavior, such as true threats, stalking, and violence.”

Finding that legislative responses were, in most circumstances, not the preferred method of addressing the problem, the Task Force determined that continued work between NGOs, academics and intermediaries would be the most meaningful way to approach the issue.

The report then offered a set of six principles for responding to Internet hate speech to guide the Working Group:

1. Create clear policies on hate speech and include them within the terms of service
2. Create mechanisms for enforcing hate speech policies
3. Establish a clear, user-friendly process for allowing users to report hate speech
4. Increase transparency about terms of service enforcement decisions

5. Actively encourage counter-speech and education to address hate speech
6. Unite industry to confront the issue of hate speech

The 2013 Report's analysis provided the inspiration for the ADL Best Practices document published in September 2014 with broad industry support. Its conclusions regarding a possible legislative response remain true today. New laws attempting to regulate online hate would be difficult to implement and difficult to enforce given the global nature of the Internet, the physical impossibility of monitoring content in real time, fundamental differences in legal systems, and ever-changing technology.

To confront cyberhate effectively, the greatest needs today are closer industry cooperation, improved voluntary enforcement of terms of service and community guidelines, greater transparency, simplified mechanisms for users to flag offensive content, and more direct interaction with stakeholders. In addition, more attention must be given to counter-speech strategies, teaching critical thinking, developing educational materials on cyberhate and raising awareness of the problem. All of these principles are reflected in the Best Practices.

For these reasons, ***we urge the Global Forum to acknowledge the important strides the industry has taken by endorsing the Best Practices (below and attached as an Appendix) and adopting them as the first part of its 2015 recommendations to the Internet industry and the broad international community of Internet users regarding hate online.*** Support from the Global Forum would underscore the international importance of these Best Practices, and call additional public attention to them around the world.

Recommended Best Practices for the Internet Industry and Internet Users

1. Providers should take reports about cyberhate seriously, mindful of the fundamental principles of free expression, human dignity, personal safety and respect for the rule of law.
2. Providers that feature user-generated content should offer users a clear explanation of their approach to evaluating and resolving reports of hateful content, highlighting their relevant terms of service.
3. Providers should offer user-friendly mechanisms and procedures for reporting hateful content.
4. Providers should respond to user reports in a timely manner.
5. Providers should enforce whatever sanctions their terms of service contemplate in a consistent and fair manner.

6. The Internet Community should work together to address the harmful consequences of online hatred.
7. The Internet Community should identify, implement and/or encourage effective strategies of counter-speech – including direct response; comedy and satire when appropriate; or simply setting the record straight.
8. The Internet Community should share knowledge and help develop educational materials and programs that encourage critical thinking in both proactive and reactive online activity.
9. The Internet Community should encourage other interested parties to help raise awareness of the problem of cyberhate and the urgent need to address it.
10. The Internet Community should welcome new thinking and new initiatives to promote a civil online environment.

Recommended Responses to Terrorist Use of Social Media

The above ten points are also important practices for responding to terrorist use of social media, but more is needed. For this reason, we provide these five additional recommendations:

1. Providers should give priority attention to how their platforms are being used by terrorists and terrorist groups to promote terrorism, to recruit potential new terrorists, and to foster self-radicalization.
2. Providers should make their expertise available to those looking to generate and promote counter-narratives.
3. Providers should work with interested stakeholders to analyze the impact of counter-narratives in terms of their reach, scope, and effectiveness.
4. Providers should consider creating a specific new terrorism category for users seeking to flag terrorism-related content.
5. Providers should use their corporate voices to condemn terrorist use of their platforms and to explain why terrorist activity and advocacy is inconsistent with their goals of connecting the world.

Underlying all of the recommendations is the understanding that rules on hate speech may be written and applied too broadly so as to encumber free expression. Thus, a underlying principle for these recommendations is that care should be taken to respect free expression and not to encumber legitimate debate and free speech.

APPENDICES

APPENDIX A: BEST PRACTICES

The following section of this report is taken from the Anti-Defamation League's website, at www.adl.org/cyberhatebestpractices. These Best Practices were inspired by the last Global Forum and the work of the Anti-Cyberhate Working Group convened by the Anti-Defamation League at the behest of Inter-Parliamentary Coalition and its Internet Task Force.

Following the Best Practices, this report includes comments from major Internet companies welcoming them. Major industry players continue to refer to them and to use them as a guidepost for improving their own response to hate online.

Cyberhate Response

BEST PRACTICES FOR RESPONDING TO CYBERHATE

Best Practices

Background

Working Group

It is our hope that the following Best Practices will provide useful and important guideposts for all those willing to join in the effort to address the challenge of cyberhate. We urge members of the Internet Community, including providers, civil society, the legal community, and academia, to express their support for this effort and to publicize their own independent efforts to counter cyberhate.

PROVIDERS

1. Providers should take reports about cyberhate seriously, mindful of the fundamental principles of free expression, human dignity, personal safety and respect for the rule of law.
2. Providers that feature user-generated content should offer users a clear explanation of their approach to evaluating and resolving reports of hateful content, highlighting their relevant terms of service.
3. Providers should offer user-friendly mechanisms and procedures for reporting hateful content.
4. Providers should respond to user reports in a timely manner.

5. Providers should enforce whatever sanctions their terms of service contemplate in a consistent and fair manner.

THE INTERNET COMMUNITY

6. The Internet Community should work together to address the harmful consequences of online hatred.

7. The Internet Community should identify, implement and/or encourage effective strategies of counter-speech – including direct response; comedy and satire when appropriate; or simply setting the record straight.

8. The Internet Community should share knowledge and help develop educational materials and programs that encourage critical thinking in both proactive and reactive online activity.

9. The Internet Community should encourage other interested parties to help raise awareness of the problem of cyberhate and the urgent need to address it.

10. The Internet Community should welcome new thinking and new initiatives to promote a civil online environment.

Best Practices

Background

Working Group

The Internet is the largest marketplace of ideas the world has ever known. It enables communications, education, entertainment and commerce on an incredible scale. The Internet has helped to empower the powerless, reunite the separated, connect the isolated, and provide new lifelines for the disabled. By facilitating communication around the globe, the Internet has been a transformative tool for information-sharing, education, human interaction, and social change. We treasure the freedom of expression that lies at its very core.

Unfortunately, while the Internet's capacity to improve the world is boundless, it also is used by some to transmit anti-Semitism, anti-Muslim bigotry, racism, homophobia, misogyny, xenophobia, and other forms of hate, prejudice and bigotry. This hate manifests itself on websites and blogs, as well as in chat rooms, social media, comment sections, and gaming. In short, hate is present in many forms on the Internet. This diminishes the Internet's core values, by creating a hostile environment and even reducing equal access to its benefits for those targeted by hatred and intimidation.

In an ideal world, people would not choose to communicate hate. But in the real world they do, all too often. And hate expressed online can lead to real-world violence, nearby or far away. The challenge is to find effective ways to confront online hate, to educate about its dangers, to encourage individuals and communities to speak out when they see it, and to find and create tools and means to deter it and to mitigate its negative impact.

Best Practices

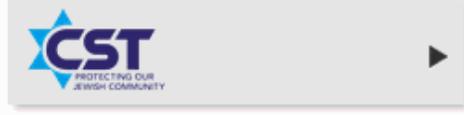
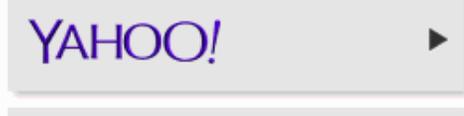
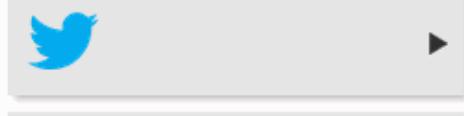
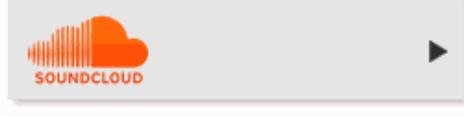
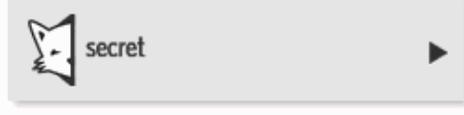
Background

Working Group

In May 2012, the Inter-Parliamentary Coalition for Combating Anti-Semitism, an organization comprised of parliamentarians from around the world working to combat resurgent anti-Semitism, asked the Anti-Defamation League (ADL) to convene a Working Group on Cyberhate. The mandate of the Working Group was to develop recommendations for the most effective responses to manifestations of hate and bigotry online. The Working Group includes representatives of the Internet industry, civil society, the legal community, and academia.

The Working Group has met four times, and its members have graciously shared their experiences and perspectives, bringing many new insights and ideas to the table. Their input and guidance have been invaluable, and are reflected in the Best Practices that we are proposing in this document. Obviously, the challenges are different for social networks, search engines, companies engaged in e-commerce, and others. Nevertheless, we believe that if adopted, these Best Practices could contribute significantly to countering cyberhate.

Industry Responses to Best Practices



Updates on technology policy issues

Fighting Online Hate Speech

Posted: Tuesday, September 23, 2014

g+ 14

Tweet 54

Like 47

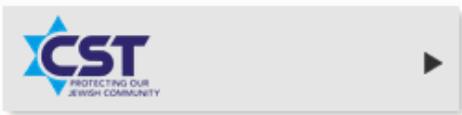
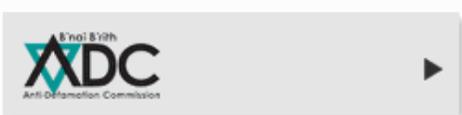
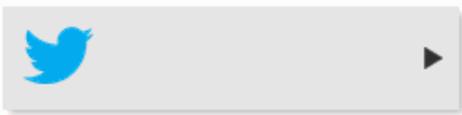
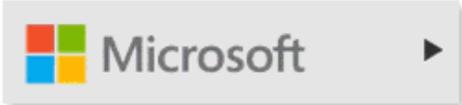
Posted by Christine Y. Chen, Senior Manager, Public Policy

Earlier today, the Anti-Defamation League (ADL) [released](#) its "[Best Practices for Responding to Cyberhate](#)." For two years, Google has participated in an industry working group convened by the ADL where, together with several other companies, other NGOs, and academics, we have exchanged insights and ideas on how to balance the need for responsible discourse with the principles of free expression. The best practices set forth by the ADL grew out of these conversations and we are excited to see them being shared with the wider Internet community.

In line with the practices set forth by the ADL, we work hard at Google to combat the spread of hateful content in order to maintain safe and vibrant communities on platforms like YouTube, Blogger, and Google+. We don't allow content that promotes or condones violence or that has the primary purpose of inciting hatred on the basis of race or ethnic origin, religion, disability, gender, sexual orientation or gender identity, age, nationality, or veteran status.

To make sure these communities stay vibrant, we also depend on our users to let us know when they see content that violates our policies. The [Google Safety Center](#) gives an overview of the tools that people can use to report content that violates our user policies on different products.





facebook.

"Facebook supports the Anti-Defamation League's efforts to address and counter cyber hate, and the Best Practices outlined today provide valuable ways for all members of the Internet community to engage on this issue," said Monika Bickert, head of global policy management at Facebook. "We are committed to creating a safe and respectful platform for everyone who uses Facebook."



Read about "**Facebook's Community Standards**" ▶



"Microsoft is committed to providing a safe and enjoyable online experience for our customers, and to enforcing policies against abuse and harassment on our online services, while continuing to keep freedom of speech and free access to information as top priorities. The Best Practices document is a tool that can foster discussion within the community and advance efforts to combat harassment and threats online."

– Dan Bross, Senior Director, Corporate Citizenship at Microsoft

- facebook. ▶
- Google ▶
- Microsoft ▶
-  secret ▶
-  SOUND CLOUD ▶
-  ▶
- YAHOO! ▶
- You Tube ▶
-  ADC
Anti-Defamation Commission ▶
-  CST
PROTECTING OUR
JEWISH COMMUNITY ▶
-  muslim advocates
PROMOTING FREEDOM & JUSTICE FOR ALL ▶

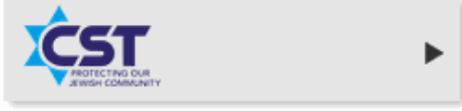
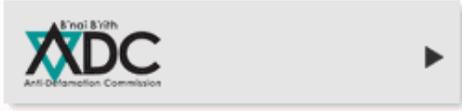
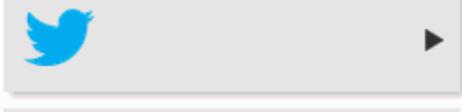
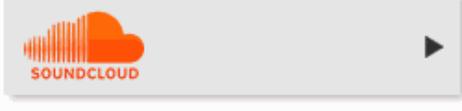
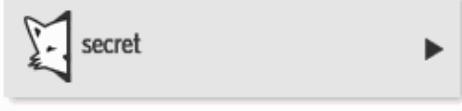


"At Secret, we work hard to create and enforce the policies needed to ensure that our community is well-lit, safe and engaging. We are committed to combating online abuse so that our users can share freely, honestly, and without judgement, as outlined in our "**Community Standards**". We support the ADL's efforts to address the challenge of hateful content online through these Best Practices."



"We are committed to providing our community with a space that is free from hatred and intolerance. Our team strives to develop and enforce policies designed to combat reported content which goes against **our Terms** and **Community Guidelines**. We fully support the ADL's continued efforts to educate and encourage individuals to take action against online hate, including the recently published Best Practice guidelines.

You can find out how to report content to our dedicated team by reading **this article**."



YAHOO!

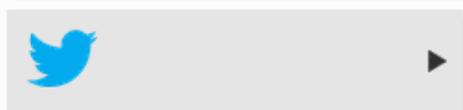
"Yahoo is committed to confronting online hate, educating our users about the dangers and realities, and encouraging our users to flag any hostile language they may see on our platform. As a member of ADL's Working Group on Cyberhate, we support the ADL's efforts to promote responsible and respectful behavior online."

Read Yahoo's Global Public Policy blog article "[Standing Up to Cyberhate](#)" ▶

You Tube



Read about YouTube "[Community Guidelines](#)" ▶



"I applaud and wholeheartedly support the incredibly important work of the ADL in fighting anti-Semitism and hatred on the Internet through Best Practices. As is clear, the Internet has become a vital recruiting tool for racists and extremists and a vehicle for them to inexpensively and easily disseminate their ideology of incitement and messages of bigotry, conspiracy theories, prejudice, bullying and calls for violence. Like the ADL, the ADC is strongly committed to meaningfully tackling this growing problem head on, and this wonderful initiative by the ADL is a critical and laudable step in promoting an inclusive, safe and respectful environment for all users.

– Dr. Dvir Abramovich, Chairman, B'nai B'rith Anti-Defamation Commission, Australia

To learn more about the [Anti-Defamation Commission](#) ►



"During the meetings between the social networks representatives and the Working Group, they came to appreciate that, for all the benefits it has brought, the internet also acts as the primary medium for encouraging and promoting hate. Their support for the statement of principles demonstrates their growing commitment to combating hate, and marks the start of a developing relationship determined to reduce harmful content which involves all the players in this process.

CST believes that this positive development will encourage users to hold the social networks to the principles contained in the statement and looks forward to continuing participation in the joint initiative."

More information on [Community Security Trust](#) ►

APPENDIX B: CYBER-SAFETY ACTION GUIDE

This Appendix features a screenshot of ADL's [Cyber-Safety Action Guide](http://www.adl.org/cybersafetyguide), which brings together in one place the relevant Terms of Service addressing hate speech of major Internet companies. That Guide is available at www.adl.org/cybersafetyguide. Individuals and groups seeking to respond to various manifestations of hate online have found it to be a unique and very useful tool, and the list of participating companies continues to grow.

Following the Guide, the Appendix includes recent statements from three major players – Facebook, Google and Twitter – regarding modifications and improvements in their process for addressing hate speech on their platforms.

ADL Cyber-Safety Action Guide

Your voice is the most powerful tool in fighting hate online.

Because of the enormous volume of content, companies typically rely on users like you to bring problems to their attention. Click on the company or product name below to quickly access its policies and a link to make your complaint heard.



[View press release ▶](#)

[Gua en español ▶](#)

We welcome your participation!

[Get updates and make suggestions ▶](#)

[amazon.com](#) ↓

[at&t](#) ↓

[comcast](#) ↓

[ebay](#) ↓

[facebook](#) ↓

[GoDaddy.com](#) ↓

[Google+](#) ↓

[Google play](#) ↓

[Instagram](#) ↓

[LinkedIn](#) ↓

[myspace](#) ↓

[PayPal](#) ↓

[Pinterest](#) ↓

ADL Cyber-Safety Action Guide

Your voice is the most powerful tool in fighting hate online.

Because of the enormous volume of content, companies typically rely on users like you to bring problems to their attention. Click on the company or product name below to quickly access its policies and a link to make your complaint heard.



[View press release ▶](#)

[Gua en español ▶](#)

We welcome your participation!

[Get updates and make suggestions ▶](#)

Continued

-  reddit ↓
-  topix ↓
-  tumblr ↓
-  ↓
-  vimeo ↓
-  YAHOO! ↓
-  YouTube ↓

APPENDIX C: REVISED INDUSTRY POLICIES AND PRACTICES

Facebook, March 15, 2015 New Community Standards Announcement

Explaining Our Community Standards and Approach to Government Requests

By [Monika Bickert](#), Head of Global Policy Management, and [Chris Sonderby](#), Deputy General Counsel

Every day, people around the world use Facebook to connect with family and friends, share information and express themselves. The conversations that happen here mirror the diversity of the more than one billion people who use Facebook, with people discussing everything from pets to politics. Our goal is to give people a place to share and connect freely and openly, in a safe and secure environment.

We have a set of [Community Standards](#) that are designed to help people understand what is acceptable to share on Facebook. These standards are designed to create an environment where people feel motivated and empowered to treat each other with empathy and respect.

Today we are providing more detail and clarity on what is and is not allowed. For example, what exactly do we mean by nudity, or what do we mean by hate speech? While our policies and standards themselves are not changing, we have heard from people that it would be helpful to provide more clarity and examples, so we are doing so with today's update.

There are also times when we may have to remove or restrict access to content because it violates a law in a particular country, even though it doesn't violate our Community Standards. We report the number of government requests to restrict content for contravening local law in our Global Government Requests Report, which we are also releasing today. We challenge requests that appear to be unreasonable or overbroad. And if a country requests that we remove content because it is illegal in that country, we will not necessarily remove it from Facebook entirely, but may restrict access to it in the country where it is illegal.

Billions of pieces of content are shared on Facebook every day. We hope these two updates help provide more clarity about the standards we have, whether they are our own Community Standards or those imposed by different laws around the world.

More Detailed Community Standards

The updated Community Standards are broken into four sections:

- **Helping to keep you safe**
- **Encouraging respectful behavior**
- **Keeping your account and personal information secure**
- **Protecting your intellectual property**

In particular, we've provided more guidance on policies related to self-injury, dangerous organizations, bullying and harassment, criminal activity, sexual violence and exploitation, nudity, hate speech, and violence and graphic content. While some of this guidance is new, it is consistent with how we've applied our standards in the past.

It's a challenge to maintain one set of standards that meets the needs of a diverse global community. For one thing, people from different backgrounds may have different ideas about what's appropriate to share — a video posted as a joke by one person might be upsetting to someone else, but it may not violate our standards.

This is particularly challenging for issues such as hate speech. Hate speech has always been banned on Facebook, and in our new Community Standards, we explain our efforts to keep our community free from this kind of abusive language. We understand that many countries have concerns about hate speech in their communities, so we regularly talk to governments, community members, academics and other experts from around the globe to ensure that we are in the best position possible to recognize and remove such speech from our community. We know that our policies won't perfectly address every piece of content, especially where we have limited context, but we evaluate reported content seriously and do our best to get it right.

If people believe Pages, profiles or individual pieces of content violate our Community Standards, they can report it to us by clicking the "Report" link at the top, right-hand corner. Our reviewers look to the person reporting the content for information about why they think the content violates our standards. People can also unfollow, block or hide content and people they don't want to see, or reach out to people who post things that they don't like or disagree with.

While the Community Standards outline Facebook's expectations when it comes to what content is or is not acceptable in our community, countries have local laws that prohibit some forms of content. In some countries, for example, it is against the law to share content regarded as being blasphemous. While blasphemy is not a violation of the Community Standards, we will still evaluate the reported content and restrict it in that country if we conclude it violates local law.

Countries contact us to let us know when content may be in violation of local laws and we compile these requests into a public report called the Global Government Requests Report.

Global Government Requests Report

The [Global Government Requests Report](#), which covers the second half of 2014, includes information about the government requests we received for content removal and account data as well as national security requests under the U.S. Foreign Intelligence Surveillance Act and through National Security Letters.

Overall, we continue to see an increase in government requests for data and content restrictions. The amount of content restricted for violating local law increased by 11% over the previous half, to 9,707 pieces of content restricted, up from 8,774. We saw a rise in content restriction requests from countries like Turkey and Russia, and declines in places like Pakistan.

The number of government requests for account data remained relatively flat, with a slight increase to 35,051 from 34,946. There was an increase in data requests from certain governments such as India, and decline in requests from countries such as the United States and Germany.

We publish this information because we want people to know the extent and nature of the requests we receive from governments and the policies we have in place to process them.

Moving forward, we will continue to scrutinize each government request and push back when we find deficiencies. We will also continue to push governments around the world to reform their surveillance practices in a way that maintains the safety and security of their people while ensuring their rights and freedoms are protected.

Twitter, February 26, 2015

Updated Safety Features

Update on user safety features

By Tina Bhatnagar ([@tinab](#)), VP, User Services

In December, we [announced](#) several product updates that were aimed at improving the safety of our users. Now we're back to tell you about the latest round of updates that are part of our long term plan.

We streamlined the process of reporting harassment on Twitter recently; now we're making similar improvements around reporting other content issues including impersonation, self-harm and the sharing of private and confidential information. These changes will begin rolling out today and should reach all users in the coming weeks.

Google, March 17, 2015

New rating system for apps

Content ratings for apps & games

To help consumers make informed choices on Google Play, we're introducing a new rating system for apps and games. These ratings provide an easy way to communicate familiar and locally relevant content ratings to your users and help improve app engagement by targeting the right audience for your content.

Note: All apps and games on Google Play are required to follow the [Google Play Developer Content Policy](#).

Starting in May, consumers worldwide will see the current Google Play rating scale replaced with their local rating on the Play Store. Territories that are not covered by a specific [International Age Rating Coalition](#) (IARC) rating authority will be assigned an age-based, generic rating.

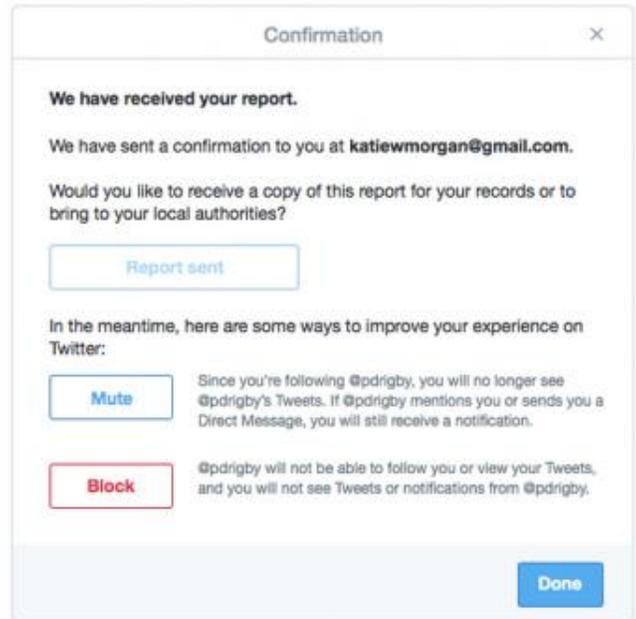
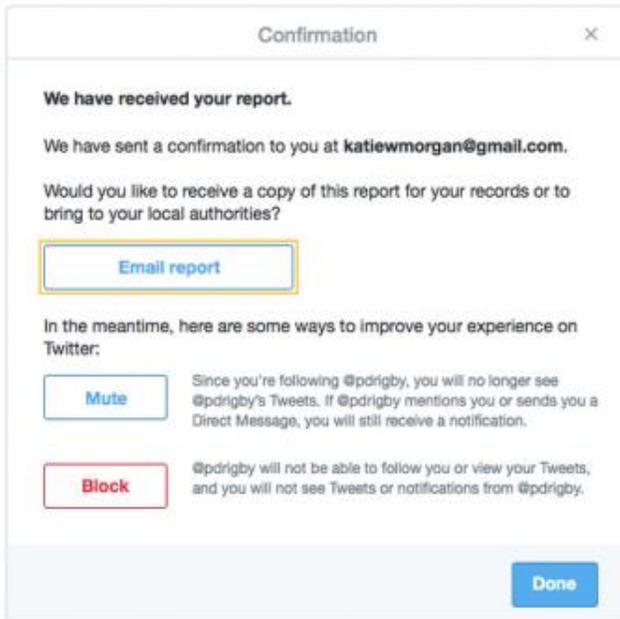
Twitter, March 17, 2015

New threat reporting tool

Making it easier to report threats to law enforcement

Today we're starting to roll out a change that makes it easier for you to report threats that you feel may warrant attention from law enforcement.

Here's how it works: after filing a report regarding a threatening Tweet directed at you, you'll see an option on the last screen to receive a summary of your report via email.



Clicking the “Email report” button will send you an email that packages the threatening Tweet and URL along with the responsible Twitter username and URL and a timestamp as well as your account information and the timestamp of your report. Our [guidelines for law enforcement](#) explain what additional information we have and how authorities can request it.



To Whom It May Concern,

We have received a report from Twitter user @katiewmorgan regarding a threat from another Twitter account, @pdrigby. The below information is a summary of the report we received.

Reported information:

Tweet: *This is a threatening tweet targeting @katiewmorgan*

URL: <https://twitter.com/pdrigby/status/123456789>

Username: @pdrigby

Account URL: <https://twitter.com/pdrigby>

Tweet sent at: 11:57 AM - 16 Mar 15

Reporter information:

Username: @katiewmorgan

Account URL: <https://twitter.com/katiewmorgan>

Report generated at: 12:07 PM - 16 Mar 15

Please refer to our Law Enforcement Guidelines (<https://support.twitter.com/articles/41949>) for guidance on how to request non-public user account information from Twitter.

Respectfully,

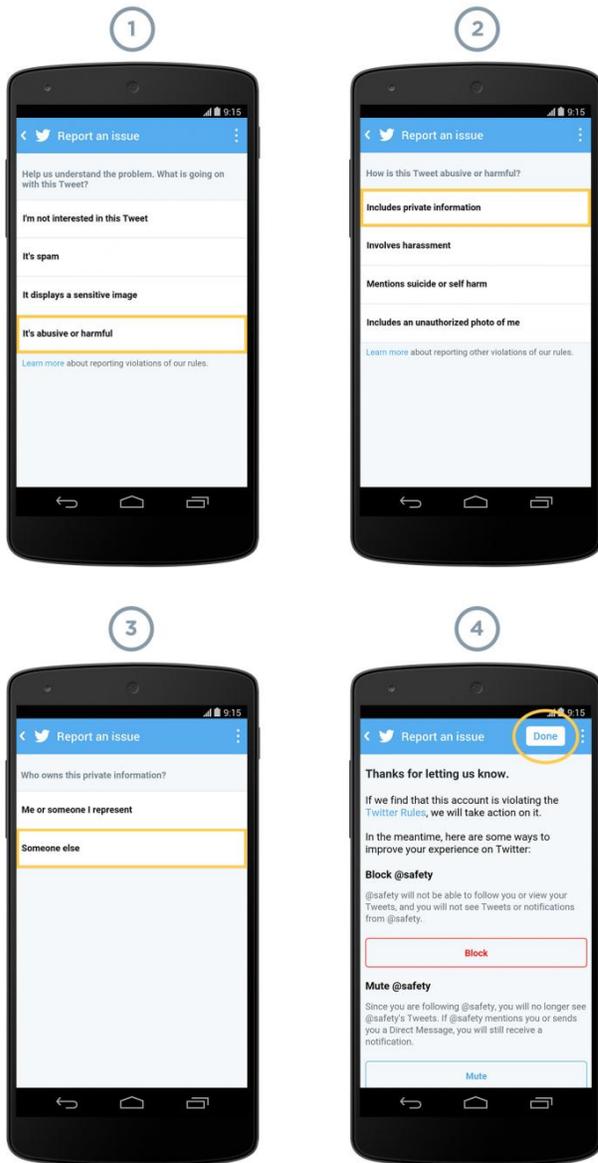
The Twitter Safety Team

[Need help?](#)

Twitter, Inc. 1355 Market Street, Suite 900 San Francisco, CA 94103

While we take threats of violence seriously and will [suspend](#) responsible accounts when appropriate, we strongly recommend contacting your local law enforcement if you're concerned about your physical safety. We hope that providing you with a summary of your report will make that process easier for you.

Finally, we'd like to acknowledge our [safety partners](#), like the [National Network to End Domestic Violence](#), for their feedback on this feature. Their input continues to be extremely valuable to us as we refine our reporting process so that it's more efficient and useful.



Over the last six months, in addition to the product changes, we have overhauled how we review user reports about abuse. As an example, allowing bystanders to report abuse – which can now be done for reports of private information and impersonation as well – involved not only an update to our in-product reporting process, but significant changes to our tools, processes and staffing behind the scenes. Overall, we now review five times as many user reports as we did previously, and we have tripled the size of the support team focused on handling abuse reports.

These investments in tools and people allow us to handle more reports of abuse with greater efficiency. So while we review many more reports than ever before, we've been

able to significantly reduce the average response time to a fraction of what it was, and we see this number continuing to drop.

We are also beginning to add several new enforcement actions for use against accounts that violate our [rules](#). These new actions will not be visible to the vast majority of rule-abiding Twitter users – but they give us new options for acting against the accounts that don't follow the rules and serve to discourage behavior that goes against our policies.

The safety of our users is extremely important to us. It's something we continue to work hard to improve. This week's changes are the latest steps in our long-term approach, and we look forward to bringing you additional developments soon.

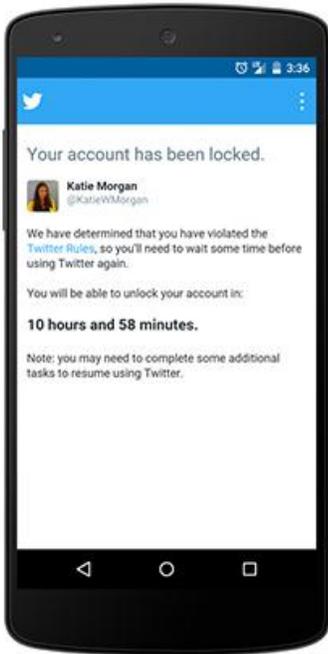
Twitter, April 21, 2015

Policy and product updates aimed at combating abuse

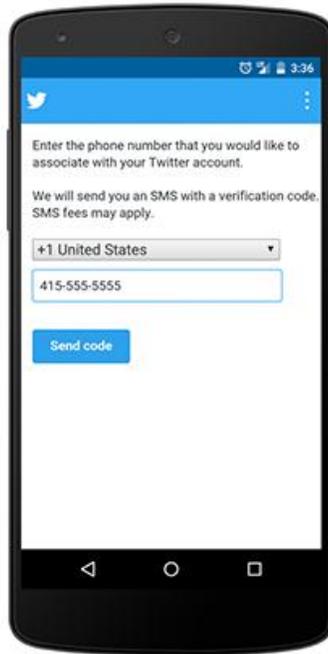
We believe that users must feel safe on Twitter in order to fully express themselves. As our General Counsel [Vijaya Gadde](#) explained last week in an opinion piece for the [Washington Post](#), we need to ensure that voices are not silenced because people are afraid to speak up. To that end, we are today announcing our latest product and policy updates that will help us in continuing to develop a platform on which users can safely engage with the world at large.

First, we are making two policy changes, one related to prohibited content, and one about how we enforce certain policy violations. We are updating our [violent threats policy](#) so that the prohibition is not limited to “direct, specific threats of violence against others” but now extends to “threats of violence against others or promot[ing] violence against others.” Our previous policy was unduly narrow and limited our ability to act on certain kinds of threatening behavior. The updated language better describes the range of prohibited content and our intention to act when users step over the line into abuse.

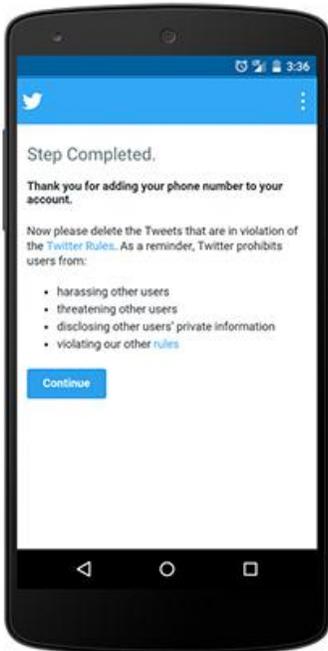
On the enforcement side, in addition to other actions we already take in response to abuse violations (such as requiring users to delete content or verify their phone number), we're introducing an additional enforcement option that gives our support team the ability to lock abusive accounts for specific periods of time. This option gives us leverage in a variety of contexts, particularly where multiple users begin harassing a particular person or group of people.



An account may be locked for a pre-defined time period.



A user may be asked to verify their phone number.



Certain actions are prohibited on Twitter.



A user may be asked to delete certain Tweets. After completing the requested actions, their account is unlocked.

Second, we have begun to test a product feature to help us identify suspected abusive Tweets and limit their reach. This feature takes into account a wide range of signals and context that frequently correlates with abuse including the age of the account itself, and the similarity of a Tweet to other content that our safety team has in the past independently determined to be abusive. It will not affect your ability to see content that you've explicitly sought out, such as Tweets from accounts you follow, but instead is designed to help us limit the potential harm of abusive content. This feature does not take into account whether the content posted or followed by a user is controversial or unpopular.

While [dedicating more resources](#) toward better responding to abuse reports is necessary and even critical, an equally important priority for us is identifying and limiting the incentives that enable and even encourage some users to engage in abuse. We'll be monitoring how these changes discourage abuse and how they help ensure the overall health of a platform that encourages everyone's participation. And as the ultimate goal is to ensure that Twitter is a safe place for the widest possible range of perspectives, we will continue to evaluate and update our approach in this critical arena.



©2015 Anti-Defamation League
605 Third Avenue
New York, NY 10158-3560
www.adl.org