Note

How a New Standard of Care Can Make Social Media Companies Better "Good Samaritans"

*Jenna Hensel**

INTRODUCTION

Social media use is on the rise and a prevalent pastime for many people. Ninety-one percent of adults who use the Internet use social media regularly; moreover, social media has become the number one activity on the Internet.[1] In fact, Instagram users upload forty million images daily, and Facebook users share 684,478 pieces of content every minute.[2] Yet, with this flux of activity, not all activity is in line with social media platforms' content moderation policies.

For example, Megan Meier is one of many who fell victim to cyberbullying.[3] After Meier had a falling out with another thirteen-year-old, Sarah Drew, Sarah's mother, Lori Drew, created a fictitious Myspace profile of a sixteen-year-old boy named Josh Evans.[4] Lori Drew flirted with Meier through her Josh Evans profile for several weeks until she abruptly switched her Josh Evans profile to become

* J.D. Candidate 2021, University of Minnesota Law School. Thank you to my family and friends for your unending love and compassion, listening ears, and constant laughter and support throughout law school and the process of writing this Note. Thank you to my partner Jonathan for everything. I am forever grateful to all of you. Thank you to Professor Rozenshtein for your guidance and expertise in helping me narrow my topic and write this Note from start to finish. Thank you to Professor Martineau for your teachings of legal research that vastly improved this Note. Last, but certainly not least, thank you to the *Minnesota Law Review* editors and staff for your extraordinary hard work and diligence in editing this Note. Finally, this Note is dedicated to anyone who has suffered abuse online. Your voices are heard, your stories matter, and you are never alone.

1. Justin P. Murphy & Adrian Fontecilla, *Social Media Evidence in Government Investigations and Criminal Proceedings: A Frontier of New Legal Issues*, 19 RICH. J.L. & TECH. 1, 1–2 (2013).

2. *Id.* at 3.

3. *See* Juliet Dee, *Cyberharassment and Cyberbullying*, *in* 2 REGULATING SOCIAL MEDIA: LEGAL AND ETHICAL CONSIDERATIONS 65, 75–77 (Susan J. Drucker & Gary Gumpert eds., 2013).

4. *Id.* at 75.

hostile towards Meier.[5] She then told Meier through her Josh Evans profile, "The world would be a better place without you."[6] Meier replied, "You're the kind of boy a girl would kill herself over," and hung herself that afternoon.[7] Despite Drew's cruelty toward Meier and breach of the Myspace Terms of Service, which prohibited harassment and providing false information, Drew suffered no consequences for her actions because no federal statute existed that regulated cyberbullying.[8]

According to Section 230 of the Communications Decency Act, social media platforms cannot be held liable for users' violations of their terms of service and are simultaneously expected to moderate objectionable content, thereby serving as Good Samaritans.[9] Yet, social media platforms are not serving as effective Good Samaritans. As a result of social media companies' current Good Samaritan content moderation practices, offensive[10] and obscene[11] user activity on social media platforms is occurring, resulting in harm to victims of users' violative

5. *Id.*

6. *Id.*

7. *Id.*

8. *Id.* at 76–77; United States v. Drew, 259 F.R.D. 449 (C.D. Cal. 2009) (granting defendant Lori Drew's motion for judgment of acquittal); *see also* Kim Zetter, *Judge Acquits Lori Drew in Cyberbullying Case, Overrules Jury*, WIRED (July 2, 2009, 3:04 PM), https://www.wired.com/2009/07/drew-court [https://perma.cc/H949-HD8N] (analyzing how Lori Drew was originally charged with four felony counts of unauthorized computer access under the Computer Fraud and Abuse Act because there was no federal statute that regulated cyberbullying, how the jury convicted her of three misdemeanor charges and deadlocked on the fourth charge, and how a federal judge acquitted Drew of her misdemeanor charges because of the vague language of the statute).

9. 47 U.S.C. § 230(c).

10. This Note will use the term "offensive" to describe social media content moderation policy violations that are less physically graphic such as postings involving bullying, harassment, catfishing, hate crimes not involving physical contact, and hate speech. An example is the violations that took place in the Megan Meier cyberbullying case. *See* Dee, *supra* note 3.

11. This Note will use the term "obscene" to describe social media content moderation policy violations that are of a more physically graphic nature such as postings involving murder, rape, terrorism, pornography, hate crimes involving physical contact, sexual assault, and physically harming someone with a weapon such as a gun or knife. An example is the case where Steve Stephens posted a video titled "Easter Day Slaughter" on Facebook of him asking a seventy-four-year-old man to say Stephens's girlfriend's name before shooting the man in the head. Kathleen Chaykowski, *Murderer's Facebook Video Sparks Manhunt, Highlights Moderation Challenges*, FORBES (Apr. 17, 2017, 4:35 AM), https://www.forbes.com/sites/kathleenchaykowski/2017/04/17/murderers-facebook-video-sparks-manhunt-highlights-apps-monitoring-challenges [https://perma.cc/D9CA-XH7T].

conduct.[12] Moreover, content moderators are making an insurmountable number of errors when monitoring content, and thus monitoring content ineffectively.[13] Clearly, social media companies' Good Samaritan content moderation strategies need to change.

This Note argues that courts should limit the standard of immunity for serving as a Good Samaritan in order to push social media companies to serve as better Good Samaritans according to Section 230.[14] These new criteria for obtaining immunity by serving as a Good Samaritan will allow social media companies to continue to uphold First Amendment values[15] and simultaneously address user harm in a more impactful way by taking a stronger stance in curing current user safety issues. Therefore, by adopting and implementing these new criteria for receiving immunity, social media companies will be performing as better Good Samaritans in line with Congress's intentions in enacting Section 230(c)(2).[16]

Specifically, these new criteria for receiving immunity by serving as a Good Samaritan will require social media companies to (1) adopt new definitions for prohibited content categories that are more objective and precise so they are able to better identify prohibited content, and thus take appropriate action on violative content; and (2) train their artificial intelligence (AI) to be proficient in these new content category definitions so that the AI can more effectively screen and remove prohibitive content.[17]

This Note explores current policies and legislation concerning social media content moderation, the results of current social media Good Samaritan content moderation practices, and how social media

12. *See infra* Parts II.B.1–2.

13. *See* John Koetsier, *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*, FORBES (June 9, 2020, 8:08 PM), https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says [https://perma.cc/H3QT-6GL9] (stating that Facebook CEO admitted that moderators made errors "in more than one out of every 10 cases"); *cf.* Marie-Helen Maras, *Social Media Platforms: Targeting the "Found Space" of Terrorists*, 21 J. INTERNET L. 3, 6–7 (2017) (describing the reactionary approach in monitoring content).

14. *See infra* Part III; 47 U.S.C. § 230(c)(2).

15. *See infra* Parts I.A.1, III.

16. *See* 47 U.S.C. § 230(b)(4) ("It is the policy of the United States . . . to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material . . . ."); Annemarie Bridy, *Remediating Social Media: A Lawyer-Conscious Approach*, 24 B.U. J. SCI. & TECH. L. 193, 212 (2018) (finding that the Good Samaritan clause was inserted into Section 230 to encourage "responsible, systematic content moderation").

17. *See infra* Part III.

companies can serve as better Good Samaritans. Part I of this Note analyzes reasons why social media companies engage in content moderation. In addition, Part I discusses Section 230 of the Communications Decency Act and relevant case law applying Section 230. Part II of this Note describes the results of current social media companies' Good Samaritan content moderation practices and why social media companies are not serving as effective Good Samaritans. Lastly, Part III of this Note outlines a suggested new set of criteria for social media companies to fulfill in order to receive immunity under Section 230 for being Good Samaritans.

I. POLICIES AND LEGISLATION FOR PROSECUTING SOCIAL MEDIA PLATFORM CONTENT MODERATION VIOLATIONS

The First Amendment states: "Congress shall make no law . . . abridging the freedom of speech . . . ."[18] The primary statute concerning speech on the Internet is Section 230 of the Communications Decency Act.[19] These sources, along with social media companies' economic motivations, influence how and to what extent social media companies engage in content moderation. Section A of this Part explains reasons why social media companies engage in content moderation, including corporate responsibility to uphold First Amendment values and preserving advertising revenue. Then, Section B of this Part discusses Section 230 of the Communications Decency Act and case law applying Section 230 to cases involving negligence and defamation claims against interactive computer service providers.[20]

A. SOCIAL MEDIA COMPANIES' MOTIVATIONS TO ENGAGE IN VOLUNTARY CONTENT MODERATION

The government does not regulate social media content moderation.[21] This means that social media companies are able to construct

18. U.S. CONST. amend. I.

19. *See* 47 U.S.C. § 230.

20. An "interactive computer service" is defined as "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions." *Id.* § 230(f)(2). Social media companies qualify as providers of interactive computer services. *See id.*

21. *See* JOHN SAMPLES, CATO INST., POL'Y ANALYSIS NO. 865, WHY THE GOVERNMENT SHOULD NOT REGULATE CONTENT MODERATION OF SOCIAL MEDIA 4 (2019) (describing how First Amendment freedom of speech protection only applies when a person is injured by an action caused by the state, and since social media platforms are not the government, social media platforms are not regulated by the First Amendment). *But see*

and implement their own policies for moderating content on their platforms.[22] This Section discusses two reasons why social media companies engage in content moderation: corporate responsibility to uphold First Amendment values and preserving advertising revenue.

1. Corporate Responsibility

Social media platforms moderate content on their websites due to a sense of corporate responsibility to uphold First Amendment values.[23] For example, in an address at Georgetown University, Mark Zuckerberg, CEO of Facebook, said that Facebook has two responsibilities related to content moderation: to remove content that could cause real danger to the best of Facebook's ability, and to uphold a wide definition of freedom of expression.[24] Yet, Zuckerberg also said that Facebook wants to allow the definition of "dangerous" to be limited to what is absolutely necessary, such as dehumanizing others through hate speech, which in turn can lead to violence.[25] Facebook's balancing of freedom of expression and user safety is exemplified through the text describing Facebook's Community Standards, which states that the goal of the Community Standards is to "create a place

VALERIE C. BRANNON, CONG. RSCH. SERV., R45650, FREE SPEECH AND THE REGULATION OF SOCIAL MEDIA CONTENT 4 (2019) ("[G]overnment regulation [of Internet content] would constitute state action that implicates the First Amendment.").

22. *See, e.g.*, *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards [https://perma.cc/TPM7-VH2B] (showing the set of rules that Facebook, a social media platform, created for its users to abide by on its website).

23. *See generally* VICTORIA L. KILLION, CONG. RSCH. SERV., IF11072, THE FIRST AMENDMENT: CATEGORIES OF SPEECH (2019) (discussing how regulations of protected speech generally receive strict or intermediate scrutiny while the government has more leeway to regulate unprotected speech). The First Amendment extends to speech "in pursuit of a wide variety of political, social, economic, educational, religious, and cultural ends." *Id.* at 1 (quoting Roberts v. U.S. Jaycees, 468 U.S. 609, 622 (1984)). Thus, speech is generally protected under the First Amendment unless it qualifies for one of the categories of unprotected speech. *Id.* Categories of unprotected speech include, among others, child pornography, speech integral to criminal conduct, and true threats. *Id.* at 1–2. The Supreme Court has ruled that speech integral to criminal conduct generally occurs when "used as an integral part of conduct in violation of a valid criminal statute." *Id.* at 2 (quoting Giboney v. Empire Storage & Ice Co., 336 U.S. 490, 498 (1949)). Additionally, the Supreme Court has ruled that true threats occur when the speaker "means to communicate a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals." *Id.* (quoting Virginia v. Black, 538 U.S. 343, 359 (2003)).

24. *Mark Zuckerberg Stands for Voice and Free Expression*, FACEBOOK (Oct. 17, 2019), https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression [https://perma.cc/XQ6D-DTDZ].

25. *Id.*

for expression and give people a voice."[26] The text goes on to say that Facebook limits expression to honor values such as safety and dignity.[27] In short, these freedom of speech initiatives are balanced against upholding values of user safety, preventing harm to users, public relations, and revenue implications for certain advertisers.[28]

Similarly, Twitter is working to increase its content moderation practices to allow free expression while protecting users from abuse. Vijaya Gadde, General Counsel of Twitter, spoke out by saying, "[f]reedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up. We need to do a better job combating abuse without chilling or silencing speech."[29] Following this statement, on December 20, 2015, Twitter published the "Twitter Rules," which are an official statement of its content moderation policies.[30] Specifically, Twitter's balancing of freedom of speech and protecting user safety is seen in the text describing the Twitter Rules.[31] This text states that "Twitter's purpose is to serve the public conversation. Violence, harassment and other similar types of behavior discourage people from expressing themselves . . . . Our rules are to ensure all people can participate in the public conversation freely and safely."[32] Thus, Twitter recently initiated a new policy for combatting dehumanizing language targeting groups of individuals.[33] Accordingly, Twitter users can report tweets that compare religions to viruses or plagues and describe groups or individuals as rodents or insects.[34] This new policy also involves Twitter's AI searching out these terms and referring the tweets at issue to a human content moderator for possible disciplinary action.[35]

26. *Community Standards*, *supra* note 22.

27. *Id.*

28. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1626 (2018).

29. Vijaya Gadde, *Twitter Executive: Here's How We're Trying to Stop Abuse While Preserving Free Speech*, WASH. POST (Apr. 16, 2015, 5:05 AM), https://www.washingtonpost.com/posteverything/wp/2015/04/16/twitter-executive-heres-how-were-trying-to-stop-abuse-while-preserving-free-speech [https://perma.cc/5LUU-4HSH].

30. *See* Klonick, *supra* note 28, at 1629.

31. *See The Twitter Rules*, TWITTER, https://help.twitter.com/en/rules-and-policies/twitter-rules [https://perma.cc/J3Z5-TK3Y].

32. *Id.*

33. Sara Harrison, *Twitter and Instagram Unveil New Ways to Combat Hate—Again*, WIRED (July 11, 2019, 7:00 AM), https://www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again [https://perma.cc/Q39H-3798].

34. *Id.*

35. *Id.*

Following this new policy, Twitter officials spoke out saying that they allow users to hide other users' content that may be objectionable but not explicitly against Twitter's Terms of Service.[36] When questioned about the policy that allows users to hide others' postings, Gadde said, "[W]e need to permit as many people in the world as possible for engaging on a public platform, and it means that we need to be open to as many viewpoints as possible."[37] Twitter officials additionally reported that as of October 2019, forty percent of suspicious content is automatically forwarded to content moderation teams.[38] Yet, despite its commitment to maintaining an open platform, Twitter recently has announced that they are banning all political ads and cause-based ads that advocate for specific outcomes.[39] Clearly, Twitter is still figuring out how to police user content while maintaining itself as an open forum for speech.

2. Preserving Advertising Revenue

Social media platforms also moderate content on their websites in order to ban material that may inhibit their advertising revenue.[40]

36. Jason Koebler & Joseph Cox, *How Twitter Sees Itself*, VICE (Oct. 7, 2019, 8:00 AM), https://www.vice.com/en_us/article/a35nbj/twitter-content-moderation [https://perma.cc/ZB65-89GL] (defining "hide" as choosing to allow the content to exist on the platform, yet not allowing it to appear on the user's account). Some critics argue that this policy "places more burden on users and more trust in software solutions . . . to police hateful or otherwise violating content on [Twitter]." *Id.*

37. *Id.*

38. *Compare id.*, *with* Harrison, *supra* note 33 (discussing how Stephanie Otway, a spokesperson of Instagram, which is owned by Facebook, stated that policing bullying is Instagram's top priority and how, in addition to using human moderators, Instagram's artificial intelligence identifies "bullying language" such as "stupid" and "ugly" and asks viewers before posting, "Are you sure you want to post this?").

39. Emily Stewart, *Twitter Is Walking into a Minefield with Its Political Ads Ban*, VOX (Nov. 15, 2019, 3:00 PM), https://www.vox.com/recode/2019/11/15/20966908/twitter-political-ad-ban-policies-issue-ads-jack-dorsey [https://perma.cc/G3FC-AVKH] (discussing the ban and concerns surrounding its implementation). Twitter is currently defining "political content" as "content that references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome." *Political Content*, TWITTER, https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html [https://perma.cc/D94H-JPZ4]; *cf. First Amendment: Freedom of Speech Political Speech*, CONST. L. REP., http://constitutionallawreporter.com/amendment-01/political-speech/#Introduction [https://perma.cc/RCS3-A6SX] (discussing how political speech is usually given the strongest protection and restrictions on political speech are judged on a strict scrutiny standard by the courts).

40. Klonick, *supra* note 28, at 1627. *See generally* Rishi Iyengar, *Here's How Big Facebook's Ad Business Really Is*, CNN BUS. (July 1, 2020, 9:19 AM), https://www.cnn.com/2020/06/30/tech/facebook-ad-business-boycott/index.html [https://perma

When users spend more time on social media websites, social media platforms' advertising revenues increase.[41] Thus, social media companies aim to create websites that match user expectations so that users choose to spend more time on their platforms.[42] Conversely, if social media websites remove too much information, users may lose trust in the websites and stop visiting the platforms.[43] In other words, social media companies ban content that violates their content moderation policies in order to attract users to actively participate on their websites, creating more advertising revenue for their websites.[44]

B. SECTION 230 OF THE COMMUNICATIONS DECENCY ACT AND RELEVANT CASE LAW

Section 230 gives interactive computer service providers such as social media platforms immunity from liability for content that users post on their websites.[45] In *Zeran v. American Online Inc.*, the court gave a compelling analysis for the two reasons Congress passed Section 230.[46] First, Congress passed Section 230 to maintain the "robust nature" of Internet communication and minimize governmental interference.[47] Second, Congress passed Section 230 to overturn the *Stratton Oakmont, Inc. v. Prodigy Services Co.* decision, and thus encourage interactive computer service providers to moderate content posted by users on their platforms and remove offensive and obscene content.[48] In the *Stratton Oakmont* case, the court held that the interactive computer service provider Prodigy, the defendant, was liable for defamatory comments made by a user under a strict liability standard.[49] This was because Prodigy acted similar to an original publisher by

.cc/AT7L-TMPN] (noting that Facebook made $69.7 billion from advertising in 2019); Lauren Feiner, *Alphabet Discloses YouTube Ad Revenues of $15.5 Billion, Cloud Revenues of $8.92 Billion for 2019*, CNBC (Feb. 3, 2020, 8:35 PM), https://www.cnbc.com/2020/02/03/alphabet-discloses-youtube-cloud-revenues-for-the-first-time.html [https://perma.cc/98RJ-CXQK] (reporting that YouTube earned $15.5 billion from advertising in 2019).

41. Klonick, *supra* note 28, at 1627.

42. *Id.*

43. *Id.*

44. *Id.*

45. 47 U.S.C. § 230.

46. Zeran v. Am. Online, Inc., 129 F.3d 327, 330–31 (4th Cir. 1997).

47. *Id.*

48. *Id.* at 331.

49. Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 031063/94, 1995 WL 323710, at *5 (N.Y. Sup. Ct. 1995).

screening and editing messages posted on its website and advertising that it monitored the content on its website.[50]

The two pertinent clauses concerning protections for interactive computer service providers such as social media platforms are Section 230(c)(1) and (c)(2).[51] Specifically, according to Section 230(c)(1) social media platforms such as Facebook and Twitter cannot be held liable by users who either publish offensive information or are affected by offensive information because social media platforms cannot "be treated as the publisher or speaker of any information provided by another information content provider."[52] This means that Section 230(c)(1) prevents social media companies from liability for hosting content where the plaintiff wants to hold the provider liable as the publisher of the content.[53]

Additionally, Section 230(c)(2) provides immunity for interactive computer service providers such as social media companies that voluntarily act in good faith to restrict access to objectionable material.[54] This subsection of Section 230 is known as the "Good Samaritan" clause because it is intended to encourage online content moderators to moderate content responsibly on their websites.[55] Section 230(c)(2) prevents social media companies from being held liable for taking good faith actions to restrict access to content that is "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable."[56] Thus, Section 230(c)(2) only immunizes interactive computer service providers' actions taken in good faith.[57] As U.S. District Judge Paul A. Magnuson said, "If the publisher's motives are irrelevant and always immunized by (c)(1), then (c)(2) is unnecessary. The Court is unwilling to read the statute in a way that renders the good-faith requirement superfluous."[58] Moreover, Facebook and Twitter currently reflect this Good Samaritan clause by advising users not to post threatening or harassing content or use their platforms for

50. *Id.*
51. 47 U.S.C. § 230.
52. *Id.* § 230(c)(1).
53. *See id.*
54. *Id.* § 230(c)(2).
55. *See supra* note 16.
56. 47 U.S.C. § 230(c)(2).
57. *See id.*
58. e-ventures Worldwide, LLC v. Google, Inc., No. 14-cv-646, 2017 U.S. Dist. LEXIS 88650, at *9 (M.D. Fla. Feb. 8, 2017).

illegal purposes.[59] Furthermore, courts have considered the Good Samaritan clause in lawsuits involving removing an app from the Google Play Store,[60] removing videos from YouTube,[61] and removing websites from Google search results.[62] Therefore, Section 230 immunizes social media companies from liability for content posted on their websites, which can be a legal and economic deterrent for social media companies to proactively take action against conduct that violates their content moderation policies.[63] Yet, Section 230 also clearly encourages social media companies to play a proactive role in moderating content to ensure an open, safe space for users.[64]

Courts have implicated Section 230 when deciding negligence and defamation claims that involve interactive computer service providers such as social media companies.[65] *Zeran* was a pivotal case in developing Section 230 because the Fourth Circuit held that interactive computer service providers are not liable for derogatory material posted through their service by third parties.[66] Additionally, the Ninth Circuit held that interactive computer service providers are only responsible for offensive conduct if they specifically encourage the development of the offensiveness of the conduct.[67] These cases

59. *See* Joshua N. Azriel, *Using Social Media as a Weapon to Harm Victims: Recent Court Cases Show a Need to Amend Section 230 of the Communications Decency Act*, 15 J. INTERNET L. 3, 5 (2011).

60. *See* Spy Phone Labs LLC v. Google, Inc., No. 15-cv-03756, 2016 U.S. Dist. LEXIS 143530, at *25–26 (N.D. Cal. Oct. 14, 2016).

61. *See* Darnaa, LLC v. Google, Inc., No. 15-cv-03221, 2016 U.S. Dist. LEXIS 152126, at *24–27 (N.D. Cal. Nov. 2, 2016).

62. *See e-ventures Worldwide*, 2017 U.S. Dist. LEXIS 88650, at *4.

63. *See* 47 U.S.C. § 230.

64. *See id.*

65. *See infra* notes 66–68, 72.

66. *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 328 (4th Cir. 1997).

67. Fair Hous. Council v. Roommates.com, LLC, 521 F.3d 1157, 1171 (9th Cir. 2008) (finding that Roommates.com used mandatory preference sections and drop-down menus where homeowner users could filter out potential co-tenants based on race, gender, sexual orientation, etc.). Roommates.com did not receive immunity under Section 230 because they encouraged potentially illegal content in violation of the Fair Housing Act and became a developer of this information through developing the discriminatory questions, answers, and search mechanisms. *Id.* Hence, the questions posed were responsible for the creation of the harmful content. *Id.*; *see also* Jones v. Dirty World Ent. Recordings, LLC, 840 F. Supp. 2d 1008, 1011–13 (E.D. Ky. 2012) (stating that the court followed the reasoning of *Roommates.com* in holding that the defendant Richie encouraged the development of the offensive content on the website through the name of the website encouraging the posting of "dirt," Richie's refusal to remove allegedly defamatory posts, and Richie's comments encouraging the offensiveness directed at the plaintiff, such as, "I love how the Dirty Army has war mentality").

demonstrate that Section 230 provides social media companies with broad protection from liability for user content on their websites.

Likewise, the Fifth Circuit held in a negligence case, *Doe v. Myspace*, that websites are not required to use age verification software in order to determine if the information on users' profiles is truthful and thus if any of their users are Internet predators.[68] In this case, a thirteen-year-old girl created a Myspace account stating she was eighteen, thereby circumventing all safety features of Myspace such as profiles of fourteen- and fifteen-year-old users automatically being set to "private."[69] In contrast to a public account, a private account limits the amount of personal information on the profile that is made available to users that are not in the user's friend network.[70] The young tween met up in person with a nineteen-year-old man she connected with on Myspace and was sexually assaulted by him.[71] In a similar manner, the Third Circuit held in a defamation case, *Green v. America Online*, that a user could not sue an interactive computer service provider when other users of the same provider gave his computer a virus and made defamatory remarks to him in a chat room because the provider had no liability under Section 230.[72] In this case, this user was harmed despite AOL's requirement that all users agree to the terms of its Member Agreement, which requires all users to conform to AOL's standards for online speech and conduct described in AOL's Community Guidelines.[73]

Ultimately, these cases demonstrate that Section 230 gives social media companies broad protections concerning third party user content on their platforms, and users are being harmed despite the current Good Samaritan content moderation controls employed by social media companies. Thus, social media companies' current content moderation practices are not satisfying Congress's expectations in enacting Section 230(c)(2), because contrary to Congress's intentions, they are not engaging in responsible content moderation.[74]

In short, Section 230 immunizes social media companies from liability for content posted by users and for good faith actions taken by social media companies to restrict objectionable content on their

68. *See* Doe v. Myspace, Inc., 528 F.3d 413, 420–22 (5th Cir. 2008).
69. *See id.* at 416.
70. *Id.*
71. *Id.*
72. *See* Green v. Am. Online, 318 F.3d 465, 470 (3d Cir. 2003).
73. *Id.* at 469.
74. *See supra* note 16.

platforms.[75] Even with this immunity, social media companies moderate content to preserve First Amendment values and maintain strong advertising revenue.[76] This phenomenon demonstrates the balance social media companies face in preserving freedom of speech and preventing user harm. However, users are still being hurt from offensive and obscene content moderation violations.[77] Social media companies have the power to combat this by refining their roles as Good Samaritans to take a more proactive stance in preventing user harm.

II. RESULTS OF CURRENT GOOD SAMARITAN CONTENT MODERATION LEGISLATION AND PRACTICES

Although Congress enacted Section 230(c)(2) to encourage social media companies to regulate objectionable content on their websites, the results of current social media Good Samaritan content moderation practices are falling short of preventing user harm.[78] Section A of this Part discusses critiques of Section 230 of the Communications Decency Act. Then, Section B examines the results of current social media Good Samaritan content moderation policies including users harming themselves due to being victims of offensive and obscene content moderation violations, and harmful sub-cultures involving rape, violence, and discrimination developing and expanding. The results of current social media Good Samaritan content moderation practices also involve content moderators reviewing content with an impermissibly high error rate.

A. CRITIQUES OF THE GOOD SAMARITAN CLAUSE

Although the Good Samaritan clause of Section 230 benefits social media companies by providing them with immunity for good faith actions to regulate objectionable material posted on their websites,[79] criticisms exist regarding the courts' interpretation of this statutory provision. For example, one critique of the Section 230 Good Samaritan clause is that *Zeran* provides incentive for interactive computer service providers, such as social media companies, not to review their postings.[80] This means that social media companies will not be held liable for not taking action against objectionable material on their

75. *See* 47 U.S.C. § 230.
76. *See supra* Parts I.A.1–2.
77. *See infra* Parts II.B.1–2.
78. *See* 47 U.S.C. § 230(b)(4).
79. *See id.* § 230(c)(2).
80. *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 328 (4th Cir. 1997).

websites.[81] Thus, if these companies are not liable for failure to review postings on their websites, they can use the money and resources that would otherwise be spent on reviewing postings for other expenditures.[82] This runs counter to Congress's goal in enacting the Good Samaritan clause of Section 230 of having social media companies moderate content posted on their platform by third parties.[83] As Professor William H. Freivogel said, "[T]he Communications Decency Act . . . fosters indecency rather than decency."[84]

Criticisms also exist regarding the real-world implications of the Section 230 Good Samaritan clause.[85] For example, another critique is that due to social media companies having little incentive to review postings, victims of online offensive and obscene content moderation violations may not have recourse against users who target them online.[86] This is because if social media companies do not monitor content due to blanket immunity, or collect identification of users who post offensive and obscene material, then the offensive and obscene material remains online, and victims cannot recover damages from violative users.[87] Therefore, new criteria are needed for obtaining immunity so that social media companies can effectively serve as Good Samaritans. While immunity from liability for content posted on their websites is available according to Section 230(c)(1),[88] social media companies can and should engage in new practices to fulfill Congress's goal in enacting the Good Samaritan clause of Section 230.[89]

B. RESULTS OF GOOD SAMARITAN CONTENT MODERATION PRACTICES

1. Individual Harm

Despite social media companies acting as Good Samaritans through enforcing content moderation policies and procedures, users are still suffering physical, psychological, and emotional harm from being targeted by other users who commit offensive and obscene

81. *See id.*

82. William H. Freivogel, *Does the Communications Decency Act Foster Indecency?*, 16 COMMC'N L. & POL'Y 17, 45 (2011).

83. *Id.*

84. *Id.*

85. *See generally* 47 U.S.C. § 230(c)(2).

86. *See* Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 117–19 (2009).

87. *Id.*

88. *See* 47 U.S.C. § 230(c)(1).

89. *See supra* note 16.

content moderation violations.[90] One example of offensive online conduct is hate speech.[91] Hate speech can be used to threaten individuals or groups of people in an attempt to initiate hatred and violence and intimidate the targeted people from participating in certain activities.[92] Victims of online hate speech suffer harmful physical and psychological effects[93] such as undergoing signs of mental distress, feeling an intensification of stigmatization that can cause increased psychological distress, and experiencing rapid pulse rate, nightmares, and post-traumatic stress disorder.[94]

In addition, social media users suffer devastating physical, psychological, and emotional harm from being targeted as victims of offensive content such as cyberbullying.[95] Adolescents who are victims of cyberbullying experience a significantly higher amount of depression than adolescents who are not victims of cyberbullying.[96] Sadly, there have been many cyberbullying cases resulting in the victim dying by suicide.[97] For example, Tyler Clementi, a freshmen at Rutgers

90. *See* Madhumita Murgia, *Facebook Leads the Way Against Cyberbullying, but Others Need to Follow*, TELEGRAPH (June 19, 2016, 4:23 PM), https://www.telegraph.co.uk/technology/2016/06/19/facebook-leads-the-way-in-online-compassion-but-others-need-to-f [https://perma.cc/4T8H-KPGD] (discussing how Facebook and Twitter are taking action to combat cyberbullying on their respective platforms).

91. *See* Johnny Holschuh, *#CIVILRIGHTSCYBERTORTS: Utilizing Torts to Combat Hate Speech in Online Social Media*, 82 U. CIN. L. REV. 953, 958 (2014); *see also Hate Speech*, WIKIPEDIA, https://en.wikipedia.org/wiki/Hate_speech [https://perma.cc/T54J-6CA7] (defining hate speech as "speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex or sexual orientation").

92. *See* Holschuh, *supra* note 91.

93. *See id.*

94. *See id.*

95. *See supra* Introduction; Dee, *supra* note 3, at 67 (defining "cyberbullying" as "any type of harassment or bullying . . . that occurs through e-mail . . . instant messaging, a website (including blogs), text messaging, videos or pictures posted on websites or sent through cellphones"). It is notable that cyberbullying can also be obscene if the cyberbullying involves physically graphic material. *See infra* note 102.

96. Melissa K. Holt & Dorothy L. Espeiage, *Cyberbullying Victimization: Associations with Other Victimization Forms and Psychological Distress*, 77 MO. L. REV. 641, 649 (2012) (finding that in the study conducted about cyberbullying related to other forms of bullying, 34.7% of cyberbullying victims reported experiencing depression compared to 14.5% of non-victims).

97. *See supra* Introduction; Dee, *supra* note 3, at 74–79 (discussing cases of cyberbullying resulting in the victim of cyberbullying dying by suicide including Phoebe Prince, an Irish emigrant who received hurtful text messages and Facebook messages where she was called an "Irish slut," a "druggie," and told that she deserved to die, and Jeffrey Johnston who was called "'creepy,' a 'stalker,' and worse"); Mary Elizabeth Gillis, *Cyberbullying on Rise in US: 12-Year-Old Was 'All-American Little Girl' Before Suicide*, FOX NEWS (Sept. 21, 2019), https://www.foxnews.com/health/cyberbullying-all

University, was cyberbullied.[98] Tyler's roommate set up a hidden webcam in their room to catch Tyler having an intimate encounter with another man.[99] Then, Tyler's roommate streamed the video live on Twitter and alerted his friends to watch it by sending them Twitter messages.[100] After discovering what his roommate had done, Tyler died by suicide.[101] Thus, the harmful effects of offensive content moderation violations include users hurting themselves or ending their lives. This can be prevented by modifying the methods social media companies use to serve as Good Samaritans in order to prevent this user harm.

Additionally, users can be harmed or choose to harm themselves due to being victims of obscene content moderation violations. An example of this type of violation is revenge pornography. Revenge pornography occurs when someone posts sexually explicit pictures of another person online without the person's consent.[102] Victims of revenge pornography have been rejected by employers, educational institutions, and potential future partners.[103] In addition, victims of revenge pornography have been stalked, harassed, and bullied.[104] Finally, victims of revenge pornography have died by suicide.[105] Thus, social media companies can improve their Good Samaritan content moderation practices to help ensure that victims of offensive and obscene content moderation violations do not harm themselves or

-american-little-girl-suicide [https://perma.cc/ZPX3-UDCS] (reporting that twelve-year-old Mallory Grossman died by suicide after being targeted on Snapchat when two girls took screenshots of their Snapchat videos depicting pictures of Mallory with the captions "Poor Mal. You have no friends" paired with laughing emojis and "You have no friends. When are you going to kill yourself?"); *Cyberbullying Pushed Texas Teen to Commit Suicide, Family Says*, CBS NEWS (Dec. 2, 2016, 10:00 AM), https://www.cbsnews.com/news/cyberbullying-pushed-texas-teen-commit-suicide-family [https://perma.cc/KX99-X73D] (describing how Brandy Vela died by suicide after receiving "abusive text messages" focused on her weight, and how classmates made dating websites featuring her picture and number, saying to call her because she was giving out free sex).

98. Dee, *supra* note 3, at 78.

99. *Id.*

100. *Id.*

101. *Id.*

102. Farrah Champagne, *The Rise of Revenge Pornography and Its Damaging Effects*, 16 CRIM. LITIG. 1, 1 (2016).

103. *Id.*

104. *Id.*; *see* Taryn Pahigan, *Ending the Revenge Porn Epidemic: The Anti-Revenge Porn Act*, 30 J.C.R. & ECON. DEV. 105, 115 (2017) (describing how Amanda Todd was persistently bullied after a photo of her naked breasts went viral and died by suicide).

105. *See* Holt & Espeiage, *supra* note 96; Pahigan, *supra* note 104.

experience harm from external sources as a result of being targeted on social media companies' platforms.

2. Societal Harm

Furthermore, another result of the existing Good Samaritan content moderation practices is that social media websites are becoming hubs for harmful sub-cultures and movements in society.[106] This Note will define "harmful sub-cultures" as individual users or groups that produce online content that explicitly violates social media content moderation rules and simultaneously perpetuates a larger toxic subculture in society such as rape culture, supremacy of a gender or race, or terrorist culture. Therefore, this Note further defines "harmful subcultures" as individual users or groups whose online activity explicitly violates content moderation rules, rather than a user or group who is known for causing harm outside of social media but posts permissible material online.

An example of a harmful sub-culture on social media is rape culture. Users' violations of content moderation rules through offensive and obscene conduct has been shown to perpetuate rape culture.[107] Perpetrators of rape are able to display information and brag about their sexual exploitations on social media.[108] Depending on the

106. *See* Jason Koebler & Joseph Cox, *The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People*, VICE (Aug. 23, 2015, 12:15 PM), https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works [https://perma.cc/3TVX-CQEM] (describing how Facebook has been attacked for facilitating gender-based harassment, the streaming of murder, and contributing to genocide, and is thus working to improve their content moderation practices to prevent these problems from occurring); David Uberti, *Facebook Wants You to Know Its Doing Something About Domestic Terrorism, Sort of*, VICE NEWS (Sept. 23, 2019, 7:10 PM), https://www.vice.com/en_us/article/ne83aq/facebook-wants-you-to-know-its-doing-something-about-domestic-terrorism-sort-of [https://perma.cc/4W3L-XBSR] (discussing that in response to a post from a Facebook user on a Christchurch attack where fifty-one people were murdered on Facebook Live, Facebook said it improved its policies and "automated detection tools" to focus on white supremacy terrorism, and has joined other technology companies to create an organization to prevent "online extremism").

107. *See* Holly Jeanine Boux & Courtenay W. Daum, *At the Intersection of Social Media and Rape Culture: How Facebook Postings, Texting and Other Personal Communications Challenge the "Real" Rape Myth in the Criminal Justice System*, 2015 U. ILL. J.L. TECH. & POL'Y 149, 169.

108. *Id.*; *see* Michael Levenson, *Facebook Video of Assault, Found by Victim's Mother, Breaks Open Case*, N.Y. TIMES (Sept. 2, 2020), https://www.nytimes.com/2020/09/02/us/Providence-sexual-assault-charges.html [https://perma.cc/R37V-SRY2] (describing how a mother discovered a video on Facebook of her sixteen-year-old daughter being sexually assaulted while unconscious by multiple men). "The sharing of the video on Facebook suggests the perpetrators were 'proud of what they had done and wanted to display this for others to see—and that in itself is disturbing.'" *Id.*

context of the use, this can rationalize the subordination of the rape victim and sexual violence as key components of rape culture.[109] In addition, people who create posts glorifying rape on social media can contribute to common components of rape culture including victim blaming, slut shaming, and masculine-aggressive sexuality.[110] In sum, as a result of social media companies not serving as effective Good Samaritans, rape culture is proliferating on social media.

Violations of content moderation policies can also be used to perpetuate sub-cultures of violence, terror, and discrimination through postings featuring violence and discrimination by the perpetrators, including terrorists. For example, terrorists use social media platforms to spread hate messages and recruit supporters to harm or kill their enemies.[111] Terrorists also use social media websites to distribute graphic and violent images, videos, and messages, such as executions, to instill fear in target populations.[112] Likewise, posts featuring violence and discrimination from non-terrorists, such as videos of beatings or killings or posts containing hate speech, can also contribute to the sub-cultures of violence and discrimination.[113] Therefore, as a result of social media companies falling short of Congress's expectations of serving as Good Samaritans,[114] victims are being hurt by harmful sub-cultures on social media websites. This Note proposes that courts should limit the standard of immunity for serving as a Good Samaritan under Section 230 in order to improve social media companies' actions as Good Samaritans.[115] The new criteria for receiving immunity

109. Boux & Daum, *supra* note 107, at 153.

110. *Id.* at 155.

111. Maras, s*upra* note 13, at 3 (utilizing the examples of al-Shabaab's recruitment for attacking the Mall of America and ISIS's recruitment of people to attack properties located on the Las Vegas Strip as examples of terrorists' activities on social media).

112. *Id.* (utilizing the examples of the May 2017 Manchester terrorist attack after an Ariana Grande concert and ISIS filming and distributing executions of target populations, including, but not limited to, journalists, government officials, and alleged spies, as examples of terrorists' activities on social media).

113. *See* Holschuh, *supra* note 91, at 953 (exemplifying hate speech through the tweet "Fucking stupid arrogant, smelly, useless, waste of life, sad excuse for a NHL hockey playing NIGGER!!!!" (quoting @GRIZZLYMARSHALL, TWITTER (Apr. 25, 2012, 7:32 PM))); Maras, s*upra* note 13, at 5 (explaining how a man filmed and posted the hanging of his eleven-month-old daughter on Facebook Live); *Mayhem & Murder: 10 Most Shocking Facebook Live Moments Ever*, ABC13 (Apr. 5, 2018), https://abc13.com/10-most-shocking-facebook-live-moments-ever-captured/3302314 [https://perma.cc/7AXQ-SQF5] (describing how a video was posted on Facebook of four men beating and kidnapping an eighteen-year-old man with special needs).

114. *See supra* note 16.

115. *See infra* Part III.

will result in reducing social media user harm, including users being hurt through the online effects of harmful sub-cultures.[116]

3. Inadequate Content Moderation Practices

Lastly, social media companies' Good Samaritan content moderation policies result in content moderators not moderating content effectively. This is particularly an issue for Facebook, one of the largest social networking sites in the world.[117] Mark Zuckerberg, co-founder and CEO of Facebook, said that billions of posts, comments, and messages exist across Facebook's servers, and as a result it is impossible to review all of them.[118] Thus, Zuckerberg said that Facebook reviews content if it is reported.[119] Facebook content moderators are asked to review more than ten million potentially violative posts per week and review all user-reported content within twenty-four hours.[120] In addition, Facebook content moderators are expected to review posts with an error rate of less than one percent.[121] Yet, due to the high amount of reported violations, Facebook is making "tens of thousands" of content moderation errors per day.[122] Thus, as a result of social media companies' Good Samaritan content moderation practices, objectionable material is being overlooked and significant errors are being made.

It is also clear that social media companies hiring more content moderators is not the cure for improving content moderation strategies. For example, in 2017 Facebook hired 10,000 more safety and security workers to review content on its platform, bringing the total to 20,000 for the year.[123] Additionally, YouTube and Google announced they would hire 10,000 more content moderators by the end of 2018, and Twitter doubled its content moderation workforce to 1,500 in 2019.[124] Yet, despite the increase in worker numbers, content

116. *See infra* Part III.

117. *See* Koebler & Cox, *supra* note 106 (noting that Facebook realized that failing to properly moderate content could harm its business earlier than other platforms).

118. Maras, *supra* note 13, at 6.

119. *Id.*

120. Koebler & Cox, *supra* note 106.

121. *Id.*

122. *Id.*

123. Elizabeth Dwoskin, Jeanne Whalen & Regine Cabato, *Content Moderators at YouTube, Facebook and Twitter See the Worst of the Web—and Suffer Silently*, WASH. POST (July 25, 2019, 12:00 AM), https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price [https://perma.cc/79VF-CCNW].

124. *Id.*

moderation issues are still occurring.[125] Thus, this Note proposes that courts limit the standard of immunity for serving as a Good Samaritan under Section 230 by requiring social media companies to adopt and enforce clearer, more precise prohibited content category definitions in order to reduce human content moderator error rates in content moderation.[126]

Facebook does use AI to detect content moderation violations with some success, but AI content moderation still has its shortcomings.[127] Facebook's AI has been very successful by detecting eighty-six percent of graphic violence-related removals, and ninety-six percent of adult nudity and sexual activity.[128] However, Facebook's AI has only detected thirty-eight percent of hate speech posts that are removed and is not very effective for posts composed in languages other than English or Portuguese.[129] Hence, Facebook's AI is not perfect in content moderation and removal because of cultural context, nuances in human language, and disagreement about what constitutes hate speech.[130]

In sum, social media companies' current Good Samaritan content moderation practices result in imperfect content moderator and AI review that is inhibiting social media companies from meeting Congress's expectations to serve as Good Samaritans.[131] Therefore, this Note proposes that courts alter the criteria for receiving immunity by acting as a Good Samaritan under Section 230 by invoking changes in content moderation to improve content moderator and AI review of objectionable material in order to help enhance social media companies' performances as Good Samaritans.[132]

III. THE NEW CRITERIA FOR SERVING AS A GOOD SAMARITAN FOR SOCIAL MEDIA COMPANIES: OVERRULING *ZERAN*

The new criteria for serving as a Good Samaritan will bolster social media companies' content moderation practices by helping them appropriately balance preserving freedom of speech and preventing user harm. Section A of this Part explains the components of the new criteria for receiving immunity by serving as a Good Samaritan. These

125. *See supra* Parts II.B.1–2; Koebler & Cox, *supra* note 106.
126. *See infra* Part III.
127. *See* Koebler & Cox, *supra* note 106.
128. *Id.*
129. *Id.*
130. *Id.*
131. *See* 47 U.S.C. § 230.
132. *See infra* Part III.

components are: (1) social media companies adopting and implementing new, more objective definitions of prohibited content categories into their content moderation practices, and (2) social media companies training their AI to proficiently utilize these definitions when screening content. Then, Section B discusses why the courts should read these criteria into Section 230 and therefore overrule *Zeran*. Next, Section C analyzes how social media content moderation practices will benefit through social media companies adopting and implementing these criteria. In addition, Section D examines how former President Trump's Executive Order on Selective Censorship highlights the need for social media companies to improve their content moderation practices. Section E then illustrates how these new criteria will prevent social media companies from being held liable under the exceptions of the Good Samaritan clause. Finally, Section F reviews and responds to counterarguments related to these criteria.

A. THE COMPONENTS OF THE NEW CRITERIA

Social media companies can fulfill Congress's intentions by effectively serving as Good Samaritans pursuant to Section 230 through adopting and enforcing new criteria for content moderation that is implemented by the courts. The new criteria for receiving immunity by acting as a Good Samaritan have two components. First, social media companies must more precisely define commonly ambiguous forms of prohibited speech such as hate speech and nudity.[133] For example, the definition of "hate speech" pursuant to the new Good Samaritan criteria is: any communication that directly attacks or threatens any person or group through pejorative language on the basis of race, ethnicity, gender, gender identity, religion, national origin, caste, sex, sexual orientation, disease that may qualify as disability under the Americans with Disabilities Act, disability, or age that incites discrimination, hostility, or violence.[134] Next, the definition of "nudity" in the new Good

133. *See* Bridy, *supra* note 16, at 220–21 (naming hate speech and nudity as ambiguous categories of commonly prohibited online material). In addition, hate speech is an example of an offensive content moderation violation and nudity is an example of an obscene content moderation violation. *See supra* notes 10–11.

134. *See* Bridy, *supra* note 16, at 220–21 (describing how hate speech is a common ambiguous category of prohibited content in social media content moderation); *cf. Hate Speech*, FACEBOOK CMTY. STANDARDS, https://www.facebook.com/communitystandards/hate_speech [https://perma.cc/PXG2-8B5S] (defining hate speech as a "direct attack . . . based on . . . protected characteristics" including race, ethnicity, religious affiliation, and caste and defining "attack" as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation"); *Hate Speech Policy*, YOUTUBE, https://support.google.com/youtube/answer/2801939 [https://perma.cc/U5SV-99WT] (defining hate speech as content

Samaritan criteria is: visible genitalia, female breasts, or buttocks except if (1) the focus is art and the genitalia, female breasts, or buttocks are not displayed in a sexually arousing manner, or (2) the focus is health, is not gratuitous, and involves breast-feeding or disseminating scientifically-based health information.[135] The second component of receiving immunity by acting as a Good Samaritan is that social media companies will train their AI to proficiently use the new content category definitions when screening content.[136]

The new prohibited speech definitions in the criteria for receiving immunity as a Good Samaritan will help content moderators correctly identify and remove prohibited content more consistently. This is because the new definitions will more precisely and objectively define the prohibited content categories as compared to the current definitions. The current definitions have room for error because they are not fully comprehensive; thus, there is a large amount of room for subjective, potentially erroneous, judgment calls.

The lack of objectivity in the current prohibited content category definitions can be seen through the current definition of "hate speech." For example, Facebook and Twitter include the category "disability" and the ambiguous category "serious disease" as protected categories in their hate speech definitions.[137] The Americans with Disabilities Act defines disability as: "[a] physical or mental impairment that substantially limits one or more major life activities, a person who has a history or record of such an impairment, or a person who is perceived by others as having such an impairment."[138] This means that certain

"promoting violence or hatred against individuals or groups" based on characteristics such as sex, gender identity and expression, age, disability, and veteran status).

135. *See* Bridy, *supra* note 16, at 220–21 (discussing how nudity is a common ambiguous category in social media content moderation and that it is hard to draw lines among art, health information, and pornography); *cf. Sensitive Media Policy*, TWITTER, https://help.twitter.com/en/rules-and-policies/media-policy [https://perma.cc/NVT4-27M3] (explaining that exceptions for adult content that is pornographic or intended to cause sexual arousal may be made for artistic, health, medical, or educational content); *Nudity and Sexual Content Policies*, YOUTUBE, https://support.google.com/youtube/answer/2802002 [https://perma.cc/C7RV-ZXJK] (explaining that YouTube allows nudity when the main purpose is educational, artistic, scientific, or documentary, and is not gratuitous).

136. *See* Koebler & Cox, *supra* note 106 (discussing how Facebook's AI has shortcomings in screening content due to disagreement about what constitutes hate speech, nuances in human language, and cultural context).

137. *Hate Speech*, *supra* note 134; *Hateful Conduct Policy*, TWITTER, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [https://perma.cc/TQ6D-TYMB].

138. *Introduction to the ADA*, ADA.GOV, https://www.ada.gov/ada_intro.htm [https://perma.cc/U9A4-97XE].

diseases can manifest into disabilities.[139] For instance, people who have HIV/AIDS,[140] multiple sclerosis,[141] or PTSD[142] often qualify as disabled under the Americans with Disabilities Act. Therefore, the new hate speech definition removes the ambiguity of "serious disease" in current social media content moderation definitions by clarifying the parameters of the severity of disease that qualifies as a protected category for hate speech. This means that content moderators will not have to make subjective judgment calls on what qualifies as a serious disease and thus will be able to adequately police more objectionable content. For these reasons, social media companies will be able to better prevent harm to users by accurately and objectively countering hate speech due to the new hate speech definition in the new prohibited speech definitions.

Further, the new prohibited speech definitions will also objectively define the category of nudity, meaning content moderators will not be called to make subjective, potentially erroneous judgment calls. For example, Facebook prohibits "fully nude close-ups of buttocks."[143] The word "close-up" is ambiguous. Rather, the new prohibited speech definition of nudity prohibits postings involving all buttocks, eliminating tricky judgment calls of defining the extent of close-ups. Likewise, Snapchat prohibits all accounts that "promote or distribute pornographic content" while simultaneously allowing "[b]reastfeeding and other depictions of nudity in certain non-sexual contexts."[144] The phrase "pornographic content" is ambiguous[145] and further clouded

139. *See infra* text accompanying notes 140–42.

140. *Fighting Discrimination Against People with HIV/AIDS*, ADA.GOV, https://www.ada.gov/hiv [https://perma.cc/42DL-57Q9] (discussing how people with the HIV/AIDS disease are protected under the Americans with Disabilities Act).

141. LAURA D. COOPER, NANCY LAW & JANE SARNOFF, ADA & PEOPLE WITH MS 1 (2019) ("The ADA covers almost everyone with MS—not only people who use wheelchairs.").

142. *See ADA: Know Your Rights, Returning Service Members with Disabilities*, ADA.GOV, https://www.ada.gov/servicemembers_adainfo.html [https://perma.cc/XG2P-ZUGH] (including PTSD as an example of a serious injury, and describing PTSD as a "hidden disability").

143. *Adult Nudity and Sexual Activity*, FACEBOOK CMTY. STANDARDS, https://www.facebook.com/communitystandards/adult_nudity_sexual_activity [https://perma.cc/CT4P-3U8M].

144. *Community Guidelines*, SNAP INC., https://www.snap.com/en-US/community-guidelines [https://perma.cc/7PZU-8HRR].

145. Taylor Kohut, *What Does "Porn" Mean Anyway?*, PSYCH. TODAY (Mar. 16, 2019), https://www.psychologytoday.com/us/blog/sex-wars/201903/what-does-porn-mean-anyway [https://perma.cc/L6P6-AZHT] (explaining the various interpretations of the meaning of pornography such as limiting pornography to the "explicit unconcealed depiction of sexual behavior" and broadening pornography to include "instances of implied nudity or sexual behavior").

by the lack of definitional clarity of the phrase "non-sexual contexts." Therefore, the new definition of "nudity" specifies the contexts that the nudity is allowed, such as if the focus is art and the nudity is not presented in a sexually arousing manner, or if the focus is health, is not gratuitous, and involves breast-feeding or disseminating scientifically-based health information. This eliminates tough judgment calls for deciding whether or not a post is pornographic and whether or not the post is presented in a non-sexual context. In short, the definitions in the new Good Samaritan criteria will help facilitate social media companies consistently and reliably serving as better Good Samaritans by clarifying exactly what content is prohibited in social media. This clarification thus eliminates the need for content moderators to make tricky, subjective judgment calls when moderating content.

B. THE COURTS AS THE IMPLEMENTERS OF THE NEW CRITERIA AND OVERRULING *ZERAN*

Courts have the authority to read these criteria for receiving immunity for serving as a Good Samaritan into Section 230 of the Communications Decency Act due to the legislative history of the Communications Decency Act. Section 230 is an amendment to the Communications Decency Act.[146] Senator James Exon introduced the Communications Decency Act because he was concerned about children being exposed to pornography on the Internet.[147] In order to gain support for the Communications Decency Act, Senator Exon created a portfolio of Internet pornography and displayed this portfolio in a blue folder on his desk where the other congressmen could observe the pornography made available to children on the Internet.[148] This folder, known as the Bluebook, was frequently cited in the debate in support of the Communications Decency Act and resulted in reluctant senators voting in support of the Communications Decency Act.[149] Thus, the Communications Decency Act was added to the Senate version of the telecommunications bill of 1995.[150] Later, in the conference committee Congress held steadily to the goal of protecting children from pornography, and as a result the Telecommunications Act of

146. *See generally* 47 U.S.C. § 230 (noting in the codification that the section was added to the Communications Act of 1934 as an amendment).

147. *See* 141 CONG. REC. S1953 (daily ed. Feb. 1, 1995) (statement of Sen. Exon).

148. *See* 141 CONG. REC. S8330 (daily ed. June 14, 1995) (statement of Sen. Exon).

149. *See, e.g.*, *id.*; 141 CONG. REC. S8089 (daily ed. June 9, 1995) (statement of Sen. Exon); 141 CONG. REC. S8332 (daily ed. June 14, 1995) (statement of Sen. Coats).

150. 141 CONG. REC. S8347, S8386–87 (daily ed. June 14, 1995).

1996 was passed including the Communications Decency Act.[151] Therefore, by serving as Good Samaritans through regulating objectionable content including pornography, social media companies are helping further Congress's intent of restricting children's access to pornography because pornography is objectionable material.[152] The proposed criteria for receiving immunity by serving as a Good Samaritan help social media companies effectively serve as Good Samaritans. In sum, the courts derive authority to read these criteria for receiving immunity for being a Good Samaritan into Section 230 from the legislative history of the Communications Decency Act because the proposed criteria are helping social media companies effectively carry out Congress's intent of protecting children from objectionable material.

Moreover, the courts are the ideal branch of government to read these criteria for receiving immunity for acting as a Good Samaritan into Section 230 rather than Congress amending Section 230. First, if Congress passed an amendment, this process would be much slower and more cumbersome than the courts reading the criteria for receiving immunity for serving as a Good Samaritan into Section 230.[153] Likewise, it is unlikely that Congress will be able to anticipate all of the situations and problems that may arise in the proposed criteria.[154] Thus, courts will be able to anticipate the situations and problems arising with these new criteria for receiving immunity for serving as a Good Samaritan more efficiently, meaning the legal change will not be unduly prolonged.[155] In addition, since the courts are more flexible than Congress, the courts can continue to adapt to the rapidly changing and developing policy area of the Section 230 Good Samaritan clause, meaning they can adapt the criteria for receiving immunity for serving as a Good Samaritan to meet the current needs of online content moderation in balancing free speech and user safety.[156]

151. *See* 142 CONG. REC. S687 (daily ed. Feb. 1, 1996).

152. *See id.*

153. *See generally* Linda D. Jellum, "*Which Is to Be Master," the Judiciary or the Legislature? When Statutory Directives Violate Separation of Powers*, 56 UCLA L. REV. 837, 862–65 (2009) (describing legislative acts in the context of separation of powers).

154. *See* Archibald Cox, *Judge Learned Hand and the Interpretation of Statutes*, 60 HARV. L. REV. 370, 370–72 (1947) (explaining how many legislators do not think about possible controversies the courts will need to settle when giving a statute meaning, and many of these controversies are unseen because they arise from consequences that Congress does not envision when enacting legislation).

155. *See id.* at 372–75.

156. *See id.* at 370–72.

Further, the courts should read these new criteria for obtaining immunity for serving as a Good Samaritan into Section 230, thereby no longer following *Zeran*. In *Zeran*, the court analyzed the plain language of Section 230, and said that the same "specter of tort liability" that discouraged social media companies from policing objectionable content also threatened an "area of . . . prolific speech" with an "obvious chilling effect."[157] However, *Zeran* is not appropriate today given the realities of what is happening to users who are victims of obscene and offensive content moderation violations. Today, users are being harmed physically, psychologically, and emotionally due to proliferation of content moderation violations on the Internet.[158] Additionally, harmful sub-cultures are developing and expanding, resulting in more users being hurt.[159] For these reasons, the policy concerns on moderating social media content should place greater emphasis on user safety similar to the emphasis placed on freedom of speech concerns. The new criteria for receiving immunity for serving as a Good Samaritan thus enable social media companies to implement heightened content moderation standards aimed at reducing harm to users. In sum, the courts should read these new criteria for receiving immunity for serving as a Good Samaritan into Section 230 and therefore overrule *Zeran* in order to empower social media companies to place greater emphasis on preventing user harm while upholding freedom of speech.

C. THE BENEFITS OF ADOPTING THE NEW CRITERIA FOR SOCIAL MEDIA CONTENT MODERATION

The new criteria for receiving immunity for serving as a Good Samaritan are consistent with Congress's objectives in enacting Section 230. First, the new criteria for receiving immunity for serving as a Good Samaritan are consistent with Congress's objective of encouraging interactive computer service providers to moderate content posted by users on their platforms and remove offensive and obscene content.[160] Additionally, the new criteria for receiving immunity for serving as a Good Samaritan satisfy Congress's goals of removing disincentives for "blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material"[161] because they empower social media companies

157. Zeran v. Am. Online, Inc., 129 F.3d 327, 331 (4th Cir. 1997).
158. *See supra* Part II.B.1.
159. *See supra* Part II.B.2.
160. *See* 141 CONG. REC. H8469–70 (daily ed. Aug. 4, 1995) (statement of Rep. Cox).
161. 47 U.S.C. § 230(b)(4).

to more consistently identify and remove prohibited content due to clarity of what qualifies as prohibited content. Lastly, the new criteria for receiving immunity for serving as a Good Samaritan are also consistent with Congress's other objective in enacting Section 230 of promoting the Internet's current development because they do not require social media companies to take any affirmative punitive action against users.[162]

The new criteria for receiving immunity for being a Good Samaritan will help social media companies continue to pursue their objectives for content moderation while meeting Congress's expectations in passing the Good Samaritan clause of Section 230. Thus, by implementing these new criteria, social media companies will continue to maintain their platforms as open forums for speech and maintain advertising revenue while rectifying the harms of current content moderation.[163] These new criteria do not require social media companies to punish users, thereby curtailing speech on their platform, or remove material that may be conducive to advertising revenues.[164] Rather, these new criteria for receiving immunity for serving as a Good Samaritan will help social media companies more effectively identify prohibited content, clarifying when it is appropriate for them to take action as Good Samaritans. Thus, social media companies will continue to self-regulate as Congress intended by regulating user conduct more effectively due to clarity of ambiguous prohibited content definitions and a more proficient AI that will screen out more potentially violative material.[165] In short, social media companies will serve as better Good Samaritans because their role as Good Samaritans will be less ambiguous, and they will continue to work on their goals for content moderation.

Furthermore, the new criteria for receiving immunity for serving as a Good Samaritan will enable social media companies to serve as more effective Good Samaritans by having them better execute this role through actively regulating offensive and obscene content moderation violations. Chief Judge Alex Kozinski said, "'One solution' to this situation . . . is that a computer service loses its Section 230 immunity 'if [it] willingly want[s] to set up not knowing who are the original content providers.'"[166] In other words, Chief Judge Kozinski

162. *See id.* § 230(b)(1).

163. *See supra* Parts I.A.1–2.

164. *See supra* Parts I.A.1–2.

165. *See* 141 CONG. REC. H8469–70 (daily ed. Aug. 4, 1995) (statement of Rep. Cox).

166. Freivogel, *supra* note 82, at 42 (citing Hon. Alex Kozinski, Remarks at the 22nd Annual Media and Law Conference, Kansas City, Mo. (Apr. 17, 2009)).

proposed that courts could implement the solution that if social media companies turn a blind eye they will not be immune from liability under Section 230.[167] Chief Judge Kozinski's solution, however, fails to take into account that Congress intended for social media companies to self-regulate objectionable content on their websites when enacting the Good Samaritan clause of Section 230.[168] The new criteria for receiving immunity for serving as a Good Samaritan purport that social media companies will be effectively taking action as Good Samaritans against content moderation violations through implementing stronger content category definitions and improving AI's content screening skills. Thus, social media companies will be fulfilling Congress's intention through the adoption and implementation of these new criteria.

D. FORMER PRESIDENT TRUMP'S EXECUTIVE ORDER ON SECTION 230

The new criteria for receiving immunity for serving as a Good Samaritan raise the bar for social media companies to receive legal immunity under the Good Samaritan clause for the purpose of preventing user harm while preserving freedom of speech. Additionally, former President Trump's Executive Order on Preventing Online Censorship also demonstrates the importance of raising the bar for social media companies to receive legal immunity and thereby simultaneously improve content moderation for the purpose of preventing selective censorship.[169] While it is currently unknown if President Joe Biden[170] will take further action on Trump's Executive Order, Biden has expressed desire to revoke Section 230 because he feels online platforms are not appropriately held liable for user content.[171] In sum, Trump's Executive Order is indicative that social media companies need to be held to higher standards for receiving legal immunity and performing content moderation.

167. *See id.*

168. *See* 141 CONG. REC. H8469–70 (daily ed. Aug. 4, 1995) (statement of Rep. Cox).

169. *See* Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (June 2, 2020).

170. Jonathan Martin & Alexander Burns, *Biden Wins Presidency, Ending Four Tumultuous Years Under Trump*, N.Y. TIMES (Nov. 7, 2020), https://www.nytimes.com/2020/11/07/us/politics/biden-election.html [https://perma.cc/J74R-X7ME] (reporting that Joe Biden was elected as the 46th President of the United States on Saturday, November 7, 2020, and defeated then-President Donald Trump in the election).

171. The Editorial Board, *Opinion: Joe Biden: Former Vice President of the United States*, N.Y. TIMES (Jan. 17, 2020), https://www.nytimes.com/interactive/2020/01/17/opinion/joe-biden-nytimes-interview.html [https://perma.cc/K3XZ-W3V8] (quoting Joe Biden as saying, "Section 230 should be revoked, immediately should be revoked, number one").

Trump's Executive Order on Preventing Online Censorship limits the legal immunity offered by Section 230 for interactive computer service providers such as social media companies.[172] Specifically, this Executive Order calls for the FCC to consider creating regulations that remove legal protection for websites such as social media companies that remove or restrict access to content that is not "obscene, lewd, lascivious, filthy, excessively violent, harassing or otherwise objectionable."[173] In short, Trump's Executive Order aims to remove blanket immunity given to social media companies by punishing social media companies for censoring material that is outside the scope of the Section 230 Good Samaritan clause.[174] While Trump's Executive Order proposes a new solution to improve social media content moderation in the area of selective censorship, the proposed criteria for receiving immunity for serving as a Good Samaritan offer a new solution to improve social media content moderation in the area of preventing user harm.[175] In short, social media content moderation needs improvement, and the new criteria for receiving immunity for serving as a Good Samaritan will help social media companies improve content moderation strategies specifically by inhibiting user harm while preserving freedom of speech.

E. THE NEW CRITERIA AND PROTECTING SOCIAL MEDIA COMPANIES FROM LIABILITY

In addition, the new criteria for receiving immunity for serving as a Good Samaritan will protect social media companies from being held liable according to the exceptions of the Good Samaritan clause. For example, social media companies can be held liable if they engage in "bad faith" by encouraging, creating, or developing offensive or illegal third-party content.[176] Following a universal set of criteria for receiving immunity will ensure across the board that social media companies are not engaging in bad faith and thus exposing themselves to liability. Thus, these new criteria will more effectively prevent social media companies from liability for encouraging or creating violative content. In addition, a second exception to the Good Samaritan clause is that social media companies are not protected from promissory

172. *See* Exec. Order No. 13,925, 85 Fed. Reg. 34,079.
173. *Id.*; *see* 47 U.S.C. § 230(c)(2)(A).
174. *See* Exec. Order No. 13,925, 85 Fed. Reg. 34,079.
175. *See id.*
176. *See supra* note 67.

estoppel, a contract suit.[177] Through implementing these new criteria, social media companies will be disincentivized from making direct promises to users because they will have the tools to perform the role of a Good Samaritan proficiently according to Congress's intent in enacting the Good Samaritan clause of Section 230.[178] In short, these new criteria for receiving immunity will more effectively protect social media companies from liability by allowing social media companies to satisfactorily serve as Good Samaritans through methods that do not put them at risk for liability.

F. COUNTERARGUMENTS ON THE NEW CRITERIA

Yet, there are valid counterarguments to these proposed criteria for receiving immunity by serving as a Good Samaritan. First, a possible counterargument is that the criteria may not be enough to enable social media companies to perform as more effective Good Samaritans. Rather, a heightened notice and takedown procedure would be more enabling for this purpose. A heightened notice and takedown procedure for social media companies would allow victims to report violations to social media companies to have the material removed which would flag potentially violative content for social media companies to review.[179] However, the problem with a heightened notice and takedown procedure is that social media companies can take down every reported post purely to align with being a Good Samaritan according to Section 230.[180] This could result in removing an overabundance of posts, thus inhibiting freedom of speech. Additionally, by the time users notify social media companies of the violative content and potentially have the violative content removed, the violative content may have spread to other sites and thus be irremovable.[181] For these reasons, victims may fear reporting.[182] Therefore, a

177. Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1109 (9th Cir. 2009) (finding that after a user contacted Yahoo! about online abuse, Yahoo! promised to put an end to the prohibited abuse but did not follow through). The court said that promissory estoppel is a contract claim that does not involve publishing and that Section 230 is irrelevant because it applies to torts involving punishing. *See id.* Additionally, the court said Yahoo! could have avoided liability by not doing anything. *See id.*

178. *See supra* note 16.

179. *See* Eric Weslander, *Murky "Development": How the Ninth Circuit Exposed Ambiguity Within the Communications Decency Act, and Why Internet Publishers Should Worry*, 48 WASHBURN L.J. 267, 297 (2008).

180. *See* Freivogel, *supra* note 82, at 46.

181. *See* Keats Citron, *supra* note 86, at 118–19.

182. *See id.* at 122–23.

heightened notice and takedown procedure would not help social media companies serve as better Good Samaritans under Section 230.

A second possible counterargument is that the new criteria for receiving immunity by serving as a Good Samaritan will be ineffective. This is because while social media companies may adopt the content category definitions into their content moderation policies and AI screening mechanisms, they may be sloppy in enforcing the new content category definitions when screening content. The government does not regulate social media content moderation, meaning that social media companies can construct and implement their own policies for content moderation.[183] Therefore, there is no mandatory requirement that social media companies adopt and enforce these new content category definitions in the proposed criteria.

Rather, social media companies receive the benefit of immunity under Section 230 for acting as a Good Samaritan by taking good faith actions to restrict access to objectionable content[184] if they choose to adopt and enforce the new criteria. This immunity is in addition to the immunity under Section 230 for content users post on their websites.[185] Thus, if social media companies choose to not simultaneously adopt and enforce the content category definitions specified in the criteria, they simply will not receive immunity for serving as a Good Samaritan if they are sued. This is the motivation for social media companies to fully enforce the new content category definitions in the proposed criteria. In short, there is no binding force compelling social media companies to simultaneously adopt and enforce the new content category definitions in the criteria for receiving immunity for serving as a Good Samaritan; however, social media companies have strong incentive to adopt and enforce the new content category definitions in the criteria to receive full immunity under Section 230.

In conclusion, the new criteria for receiving immunity for serving as a Good Samaritan that is implemented by the courts involve social media companies adopting and implementing new prohibited content category definitions into their human and AI content moderation practices. Social media companies will be able to maintain freedom of speech and more effectively target preventing user harm due to the new, clear, and objective definitions detailing what content is enjoined on social media websites. In addition, social media companies will be able to better satisfy Congress's objective in enacting Section 230 of

183. *See supra* Part I.A.
184. 47 U.S.C. § 230(c)(2).
185. *Id.* § 230(c)(1).

encouraging responsible and systematic content moderation[186] through self-regulating objectionable content more effectively. Lastly, the new criteria will help shield social media companies from liability. For these reasons, the new criteria for receiving immunity will cure current pitfalls in social media content moderation effectively and efficiently.

CONCLUSION

Social media companies are immune from liability for content that users post on their websites due to Section 230, which provides that social media companies are not the publisher or speaker of anything users post. Yet, with this immunity comes the expectation that social media companies serve as Good Samaritans by self-regulating objectionable content. In light of these concerns, social media companies moderate content to uphold freedom of speech and preserve advertising revenue. Yet, social media companies' current Good Samaritan content moderation practices are failing. Despite current social media Good Samaritan content moderation practices, users are experiencing harm, harmful sub-cultures are proliferating, and content moderators are making significant errors when moderating content.

This Note argues that in order for social media companies to serve as better Good Samaritans and thus meet Congress's expectations, the courts should read into Section 230 new criteria for receiving immunity by serving as a Good Samaritan, which involve social media companies adopting more precisely defined and objective prohibited content definitions. These new criteria also entail that social media companies train their AI to screen content according to these definitions. These new criteria will help social media companies meet Congress's Good Samaritan expectations, avoid liability for the Good Samaritan clause exceptions, and continue to regulate content to preserve advertising revenue and freedom of speech values. Freedom of speech continues to be a salient value in social media. However, due to the vast amount of user harm occurring, user harm needs to be treated as a more significant priority on a similar level as freedom of speech. These new criteria for receiving immunity by serving as a Good Samaritan will allow social media companies to effectively balance preserving freedom of speech and curtailing user harm, meeting their needs and their users' needs.

186. *See supra* note 16.