I SAW IT ON FACEBOOK SO IT MUST BE TRUE: THE EFFECTS OF MISINFORMATION TAGS

Alexandra VanBlaricum[*]

TABLE OF CONTENTS

* J.D. Candidate, University of Alabama School of 2023; B.A. Louisiana Tech University 2018.
I would like to thank my family and friends for both their unending support and unwavering loyalty.

I. THE MISINFORMATION PROBLEM

Misinformation, defined as inaccurate or misleading information,[1] spreads rapidly on Internet platforms.[2] On these platforms, users can post and access content without any "external filters" changing the form or accessibility of public information.[3] The direct link between producers and consumers of information on these platforms allows users to more readily access information.[4] This accessibility to information fostered the growth and spread of misinformation on the Internet and even led the World Economic Forum to list mass misinformation "as one of the main threats for the modern society."[5]

During the COVID-19 pandemic, social media platforms, like Twitter, Instagram, and Facebook, created policies targeting misinformation about COVID-19.[6] Section 230 of the Communications Decency Act of 1996 allows these Internet entities to censor posts on their sites.[7] While Section 230 currently protects these misinformation policies, both political parties are fighting to either amend or repeal

1. *Misinformation and Disinformation: Thinking Critically About Information Sources*, CSI LIBRARY, https://library.csi.cuny.edu/misinformation (last visited Nov. 27, 2021) (defining misinformation as the "sharing of inaccurate and misleading information in an unintentional way").

2. Alessandro Bessi & Walter Quattrociocchi, *Disintermediation: Digital Wildfires in the Age of Misinformation*, 86 AUSTL. Q. 34, 34 (2015) (discussing the "virality" of misinformation on the Internet). While the cited article focuses more generally on all Internet platforms, this article mostly focuses on misinformation on social media platforms like Facebook, Twitter, and Instagram.

3. *Id.* at 35. News sources act as external filters of information. By choosing what to report, news sources determine what the public considers important. Social media content lacks these "external filters" since each user can post the content he or she finds important, and that information is directly accessible by other users. These users do not rely on a news source to discover, verify, or share information.

4. *Id.* The accessibility has changed the way "users are informed, debate, and form their opinions." *Id.*

5. *Id.* The ability of people to easily access and spread information has led some theorists to speculate about the existence of collective intelligence. *Id.* Collective intelligence is the "production of knowledge emerging from the synergy of people working and discussing together, on the web, without physical barriers." *Id.* The World Economic Forum viewed this "unfiltered information flow" as a major threat akin to terrorism. *Id.* Without filters, people can easily find information with which they agree and share that information to their friends and other like-minded individuals. *Id.* This sharing of information can confuse people and lead to "speculation, rumours, and mistrust" that harm the public. *Id.*

6. *See* Jeff Tollefson, *The Race to Curb the Spread of COVID Vaccine Disinformation*, NATURE (Apr. 16, 2021), https://www.nature.com/articles/d41586-021-00997-x (describing Twitter's and Facebook's misinformation policies); Ben Gilbert, *Instagram Is Targeting Fake Coronavirus News and Finally Taking Disinformation and Hoaxes Seriously*, INSIDER (Mar. 24, 2020, 9:58 AM), https://www.businessinsider.com/instagram-changes-moderation-policy-for-coronavirus-hoaxes-2020-3 (describing Instagram's misinformation policies).

7. 47 U.S.C.A. § 230(c)(2) (West) (allowing Internet service providers to moderate content that is considered "obscene, lewd, lascivious, filthy, excessively violent, harassing or otherwise objectionable, whether or not such material is constitutionally protected.").

this provision.[8] Further, some researchers argue that these social media policies may be causing more harm than good.[9]

This article explores the reasons why the current social media misinformation policies are ineffective solutions to the misinformation problem. Section II explains how misinformation continues to spread on social media platforms despite these platforms' attempts to curtail misinformation on their sites. Section III examines the ineffectiveness of misinformation tags and the psychological ramifications of using these tags. Section IV explores the responses from the executive and legislative branches, state lawmakers, and the public regarding both Section 230 and social media misinformation policies. Ultimately, this article aims to show that current social media policies are not only ineffective but also contribute to the misinformation problem.

II. ALGORITHIMS AT WORK: THE SPREAD OF MISINFORMATION ON SOCIAL MEDIA

Many Americans rely on social media platforms as a main source of information.[10] When enough people believe misinformation, that misinformation "may form the basis for political and societal decisions that run counter to a society's best interests."[11] Consequently, the COVID-19 pandemic incentivized social media platforms to create stronger misinformation policies.[12] Social media platforms address misinformation in different ways.

---

8. Meghan Anand et al., *All the Ways Congress Wants to Change Section 230*, SLATE (Mar. 23, 2021, 5:45 AM), https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html. Republicans have criticized the provision because it allows social media companies to remove content. *Id.* These lawmakers argue that the policies disproportionately apply to conservative content. *Id.* Democratic lawmakers argue that the provision protects Internet platforms that do not remove harmful content from their sites. *Id.*

9. *See generally* Jason Schwartz, *Tagging Fake News on Facebook Doesn't Work, Study Says*, POLITICO (Sept. 11, 2017, 6:20 PM), https://www.politico.com/story/2017/09/11/facebook-fake-news-fact-checks-242567 (describing a study on the ineffectiveness of misinformation tags).

10. Derek E. Bambauer, *How Section 230 Reform Endangers Internet Free Speech*, BROOKINGS (July 1, 2020), https://www.brookings.edu/techstream/how-section-230-reform-endangers-internet-free-speech/ (arguing that the use of social media platforms today is comparable to the use of newspapers as a source of information in the 20th century).

11. Stephan Lewandowsky et al., *Misinformation and Its Correction: Continued Influence and Successful Debiasing*, 13 PSYCH. SCI. PUB. INT. 106, 107 (2012).

12. Barbara Ortutay & Amanda Seitz, *Defying Rules, Anti-Vaccine Accounts Thrive on Social Media*, AP NEWS (Mar. 11, 2021), https://apnews.com/article/anti-vaccine-accounts-thrive-social-media-e796aaf1ce32d02e215d3b2021a33599 (explaining that these social media platforms have pledged to fight misinformation).

Facebook's misinformation policy targets both individual users and pages.[13] Facebook tags individual pages that repeatedly share flagged information and warns users who go to those pages.[14] Additionally, Facebook's algorithm limits the spread of a user's posts when that individual repeatedly shares misinformation.[15] Facebook also reduces a post's "reach" once it has been flagged.[16] Finally, Facebook informs a user if he or she has shared a post that a fact-checker later flagged as misinformation.[17]

Instagram adopted similar COVID-19 misinformation guidelines in 2020.[18] Instagram removed COVID-19 accounts from the recommendations page, and only COVID-19 posts by reputable health organizations appear on the Explore page.[19] The platform also downranks flagged content so that it appears less frequently in users' feeds or stories.[20] Finally, at the suggestion of health authorities, Instagram removes posts that can potentially harm users.[21]

Twitter also adds misinformation tags to user posts.[22] Twitter's tags not only warn users that content is potentially misleading, but also provide links to public health resources.[23] Twitter also implemented a strike system for users who violate the platform's pandemic rules.[24] These rules ban tweets related to false treatment methods, prevention claims, transmission claims, and vaccination information

13. Pranav Mukul, *Explained: How New Facebook Feature Flags Misinformation*, INDIAN EXPRESS (May 28, 2021, 7:34 AM), https://indianexpress.com/article/explained/facebook-misinformation-fake-news-tool-7332659/.

14. *Id.* Users who visit these flagged pages receive a prompt that says, "This Page has repeatedly shared false information." *Id.* The prompt also lets the user click a link to learn more about both the reason for the flag and Facebook's fact-checking process. *Id.*

15. *Id.* A flagged user's posts will not be distributed to as many users' News Feeds. *Id.*

16. *Id.* The reach of a post relates to the number of users who will see that post in their feeds. *Id.*

17. *Id.* This notification also contains a link to the article a fact-checker read before flagging the post as false and contains a prompt that allows the user to share the verified article. *Id.*

18. Gilbert, *supra* note 6. Facebook is Instagram's parent company, so their policies are very similar, and Instagram tried to model its policies to better match Facebook's. *Id.*

19. *Id.*

20. *Id.*

21. *Id.* Posts that can harm users may include misinformation about vaccinations or COVID-19 transmission and treatment. *Id.*

22. Taylor Hatmaker, *Twitter Rolls Out Vaccine Misinformation Warning Labels and a Strike-Based System for Violations*, TECH CRUNCH (Mar. 1, 2021, 2:42 PM), https://techcrunch.com/2021/03/01/twitter-vaccine-misinformation-labels-strike-system/. Twitter stated that the goal of these policies was to limit the spread of misinformation that could affect the distribution and use of COVID-19 vaccinations. *Id.*

23. *Id.* These resources provide the users with accurate information that refutes the original misinformation. *Id.*

24. *Id.*

relating to COVID-19.[25] If a user gets "two or three strikes," his or her account may be locked for twelve hours.[26] After four strikes, the user is locked out of his or her account for one week.[27] Once a user gets five strikes, Twitter can permanently suspend the user's account.[28] Twitter also bans users for sharing misinformation about elections or voting and has developed policies to label misinforming tweets and reduce their reach.[29]

Despite these policies, misinformation still spreads on social media.[30] Because of the sheer amount of data involved, social media platforms rely on algorithms to regulate user content.[31] These algorithms misflag information and contribute to the spread of misinformation.[32] People are simply more likely to engage with content with which they agree.[33] As this content gets more engagement, social media algorithms boost the content's relevancy and show it to more users whose own biases will prompt even more engagement.[34] These algorithms have led some critics to argue that such "business models . . . encouraged the platforms to serve up

25. *Id.*

26. *Id.*

27. *Id.*

28. *Id.*

29. Todd Spangler, *Twitter Is Asking Users to Flag Misinformation, Including About COVID and Elections*, VARIETY (Aug. 17, 2021, 1:46 PM), https://variety.com/2021/digital/news/twitter-users-flag-misinformation-covid-elections-1235043215/. In August of 2021, Twitter also began testing allowing users to flag content as misleading. *Id.*

30. Ortutay & Seitz, *supra* note 12 (describing the amount of misinformation still accessible on these platforms).

31. Bambauer, *supra* note 10. These platforms largely suffer from a "scale of . . . data" problem because the platforms simply have too much data to moderate content effectively. *Id.* "[O]ver 1.7 billion" people use Facebook every day, and "half a billion" tweets are posted on Twitter daily. *Id.* Only a "fraction of posts with false or unverified information are checked and marked as such." Mark Weinstein, *Why Social Media Censorship is Worse Than Useless*, N.Y. POST (May 22, 2020, 8:34 PM), https://nypost.com/2020/05/22/why-social-media-censorship-is-worse-than-useless. Furthermore, the algorithms that platforms use to combat the sheer amount of data involved are often not perfect and can misflag content. Bambauer, *supra* note 10.

32. Chris Meserole, *How Misinformation Spreads on Social Media—And What to Do About It*, BROOKINGS (May 9, 2018), https://www.brookings.edu/blog/order-from-chaos/2018/05/09/how-misinformation-spreads-on-social-media-and-what-to-do-about-it/ (explaining how algorithms boost content that users engage with and that these boosted posts may contain misinformation); *see also* Ortutay & Seitz, *supra* note 12 (arguing that Instagram's algorithms "cross-pollinate[s] misinformation" by showing users related conspiracy theory pages and posts); Gilbert, *supra* note 6 (describing an incident where Facebook's algorithm mistakenly removed correct COVID-19 information after identifying it as misinformation).

33. Meserole, *supra* note 32 (arguing that humans are "more likely to react to content that taps into our existing grievances and beliefs"); *see also* Bessi & Quattrociocchi, *supra* note 2, at 36 (arguing that confirmation bias can contribute to the virality of information).

34. Meserole, *supra* note 32. Algorithms also tend to boost posts with inflammatory or provocative misinformation since users generally engage more with such posts as compared to posts with correct, yet less evocative, information. *Id.*

engaging, if false, coronavirus misinformation in order to profit from advertising."[35] In other words, because these posts receive more engagement, social media platforms are incentivized to allow these posts to spread so that the platform can generate profits.[36]

In 2018, after a violent attack on pedestrians in Toronto, reporter Natasha Fatah tweeted two different eyewitness accounts.[37] These two tweets show the social media relevancy algorithm at work.[38] Her first tweet described the attacker as Middle Eastern, and the other tweet, posted shortly after, correctly identified the attacker as a white man.[39] Five hours after the initial tweets, the Middle Eastern eyewitness account "received far more engagement," and twenty-four hours later, users still engaged more with the incorrect tweet.[40] The graphic below compares the performance of both tweets over twenty-four hours.

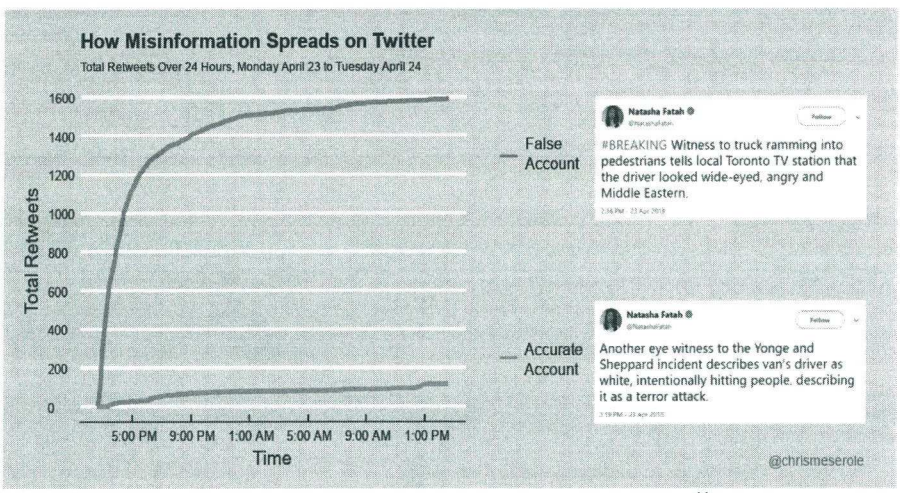


Figure 1. How Misinformation Spreads on Twitter.[41]

35. Ortutay & Seitz, *supra* note 12.
36. *Id.*
37. Meserole, *supra* note 32.
38. *Id.*
39. *Id.*
40. *Id.*
41. Chris Meserole, *How Misinformation Spreads on Twitter* (illustration), *in* Meserole, *supra* note 32. Even though the correct and incorrect tweets were posted at roughly the same time, more people shared the incorrect tweet. Meserole, *supra* note 32. Meserole notes that a potential reason why this information was shared more by users is because of its inflammatory nature. *Id.* As the initial audience reacted to the tweet, they shared it, and the algorithm then boosted the tweet to a larger audience, resulting in a broader spread of the misinformation. *Id.*

These social media algorithms also contribute to a user's confirmation bias.[42] In July 2021, U.S. Surgeon General Dr. Vivek Murthy argued that social media "algorithms tend to give [users] more of what [they] click on, pulling [them] deeper and deeper into a well of misinformation."[43] The sheer amount of information available on social media platforms makes it possible for users to find information with which they agree.[44] As users engage with that information, the algorithm will show them similar sources that align with the initial post.[45]

A study conducted by the Center for Countering Digital Hate examined this relevancy algorithm.[46] The researchers created fifteen profiles, and these profiles followed different types of groups.[47] For example, an account would follow either credible health organizations, vaccine conspiracy pages, or a mixture of both.[48] The study found that "[p]rofiles following wellness influencers and vaccine opponents were served up posts with false claims about COVID-19" and other forms of misinformation.[49] Profiles that followed only credible health organizations received no misinformation.[50] The posts containing misinformation appeared in both the suggested posts in the user's feed and on the user's Explore page.[51] This confirmation bias—assisted by social media algorithms—creates public confusion and contributes to the virality and strength of misinformation.[52]

---

42. Bessi & Quattrociocchi, *supra* note 2, at 35–36. On social media platforms, people can find information with which they agree. *Id.* at 35. This information serves to confirm their biases, making those beliefs stronger. *See id.* At its most dangerous level, widespread confirmation bias can "influenc[e] the public perception of reality on very relevant issues such as health or political economy." *Id.* at 36.

43. Todd Spangler, *Biden: Platforms Like Facebook Are 'Killing People' with Vaccine Misinformation,* VARIETY (July 16, 2021, 1:34 PM), https://variety.com/2021/digital/news/biden-facebook-vaccine-misinformation-killing-people-1235021915/.

44. Bessi & Quattrociocchi, *supra* note 2, at 35.

45. Spangler, *supra* note 43.

46. Shannon Bond, *Instagram Suggested Posts to Users. It Served Up COVID-19 Falsehoods, Study Finds,* NPR (Mar. 9, 2021, 12:01 AM), https://www.npr.org/2021/03/09/975032249/instagram-suggested-posts-to-users-it-served-up-covid-19-falsehoods-study-finds (describing a study on Instagram's social media algorithms and how those algorithms contributed to the spread of misinformation).

47. *Id.*

48. *Id.*

49. *Id.* Even when these profiles "also followed credible health organizations," Instagram's algorithms still recommended pages that regularly post misinformation. *Id.*

50. *Id.*

51. *Id.* The Explore page contains curated posts that Instagram's algorithms believe a user may find interesting. *Id.*

52. Bessi & Quattrociocchi, *supra* note 2, at 36 (arguing that phenomena such as confirmation bias and disintermediation contribute to the virality of information); *see also* Meserole, *supra* note 32

Misinformation tags also facilitate the spread of misinformation.[53] One study shows that while a label does limit the spread of a particular misinforming post, labeled tweets generally reach more users than tweets that are not labeled.[54] Users also tended to post the labeled tweets more often on other social media platforms like Facebook, Instagram, and Reddit.[55] Thus, these labels are a "superficial solution that fails to address [the] structural flaw" of most social media platforms.[56]

Even though social media platforms created stronger misinformation policies during the COVID-19 pandemic, misinformation continues to spread on their platforms.[57] The sheer amount of information on social media platforms forces the platforms to rely on algorithms.[58] These same algorithms also try to boost engagement on the platform by showing users the content they want to see.[59] Because some users generally engage with misinformation, the very algorithms used to combat misinformation continue to misinform these users.[60] In addition to these faulty algorithms, misinformation tags also contribute to misinformation reaching a broader audience.[61] Thus, the two principal elements of current social media policies—algorithms and misinformation tags—cannot successfully stop the spread of misinformation.[62]

(arguing that algorithms "can turn social media into a kind of confirmation bias machine . . . perfectly tailored for the spread of misinformation").

53. James Devitt, *Despite Warning Labels, Trump's Election Misinformation Tweets Spread Widely Across Social Media Platforms, New Study Finds,* NYU (Aug. 24, 2021), https://www.nyu.edu/about/news-publications/news/2021/august/despite-warning-labels—trump-s-election-misinformation-tweets-s.html (describing a study that found certain flagged tweets spread more than unflagged posts).

54. *Id.* The study looked at 1,149 tweets from President Trump. *Id.* The tweets included those with no tags, those with potentially misleading tags, and those with misinformation tags (and blocked engagement on Twitter). *Id.* They found that the potentially misleading tweets "spread further" on Twitter than the untagged posts. *Id.* The blocked tweets (with misinformation warnings) spread "longer and further" on other social media platforms than potentially misleading tweets and tweets with no tags. *Id.*

55. *Id.* These cross-platform posts also received "more visibility." *Id.*

56. Mark Weinstein, *Why Social Media Censorship is Worse Than Useless*, N.Y. POST (May 22, 2020), https://nypost.com/2020/05/22/why-social-media-censorship-is-worse-than-useless/.

57. Ortutay & Seitz, *supra* note 12 (describing both the ineffectiveness of current social media policies and the amount of misinformation still accessible on these platforms).

58. Bambauer, *supra* note 10.

59. Spangler, *supra* note 43.

60. *Id.*

61. Devitt, *supra* note 53.

62. *See generally* Mukul, *supra* note 13; Gilbert, *supra* note 6; Hatmaker, *supra* note 22 (describing Facebook's, Instagram's, and Twitter's social media policies); *see also* Ortutay & Seitz, *supra* note 12.

III. AN INEFFECTIVE SOLUTION: THE PSYCHOLOGICAL IMPLICATIONS OF MISINFORMATION TAGS

The spread of misinformation about COVID-19 vaccinations prompted researchers to look at the effect of misinformation and how such misinformation spreads.[63] Most recent research on misinformation focuses on COVID-19 and the 2020 presidential election because misinformation in these areas can actively harm individuals and the public.[64] While the studies described below are by no means exhaustive, misinformation tags do appear to make it slightly more likely that people can correctly judge misinformation.[65] However, misinformation tags also contribute to the "backfire effect"[66] and the "implied truth effect,"[67] creating different misinformation problems altogether.

A. The Backfire Effect

One study found a "backfire effect," a phenomenon resulting from some people's preference to believe misinformation that comports with his or her own beliefs.[68] The backfire effect is a "process by which people implicitly counterargue against any information that challenges their worldview."[69] When this effect is at work, people tend to resist corrections.[70] Retractions can also backfire and make the individual's belief in that misinformation stronger.[71]

In this study, participants accepted information that comported with their beliefs but were critical of information that countered those beliefs.[72] A person experiencing this backfire effect defends his or her opinions by relying on the source of those beliefs.[73] Because opposing information will not fit within the framework of that person's beliefs, the person will find that "no opposing information is . . . sufficiently credible to overturn dearly held prior knowledge."[74]

---

63. Tollefson, *supra* note 6.

64. *Id.*

65. Schwartz, *supra* note 9 (finding that the tags made it only 3.7% more likely that participants would correctly identify misinformation).

66. *See generally* Lewandowsky et al., *supra* note 11, at 119 (describing the backfire effect).

67. Gordon Pennycook et al., *The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings*, 66 MGMT. SCI. 4944, 4945 (2020) (describing the implied truth effect).

68. Lewandowsky et al., *supra* note 11, at 119.

69. *Id.*

70. *Id.*

71. *Id.*

72. *Id.*

73. *See id.*

74. *Id.*

Thus, in some instances, misinformation tags can backfire and increase an individual's belief in false information.[75] A similar effect can also be associated with conspiracy-minded individuals because banning content proves in their minds the value and veracity of that information.[76] However, subsequent studies have not been able to confirm whether this backfire effect exists at all.[77]

B. The Implied Truth Effect

The "implied truth effect" is a phenomenon that results from the lack of misinformation tags on some posts.[78] When some posts are flagged but not others, users perceive the unflagged posts as credible.[79] Social media platforms largely rely on misinformation tags to combat the spread of misinformation.[80] But the sheer number of posts makes it impossible to verify the accuracy of each one.[81] Thus, the benefit of tags may be outweighed by the "aura of legitimacy" the tags give unflagged misinformation.[82]

Further, misinformation is far easier to create than it is to debunk.[83] Users are also learning how to phrase posts to avoid misinformation tags.[84] Because unflagged misinformation will continue to be a problem on social media platforms, the implied truth effect will continue to influence users.[85] One potential way to address this implied truth effect is by placing "verified labels" on content instead of misinformation tags.[86] One study found that these verified labels countered the

---

75. Pennycook et al., *supra* note 67, at 4945.

76. Weinstein, *supra* note 56 (arguing that censorship is a "counterproductive" measure for refuting conspiracy theories). Censorship leads conspiracy theorists to believe that the censored information is important merely *because* the relevant authorities are trying to hide that information. *See id.*

77. Pennycook et al., *supra* note 67, at 4955 (explaining that the backfire effect is "quite elusive").

78. *Id.*

79. *Id.* at 4954; *see also* Nicolás Rivero, *The Risk of Putting Warning Labels on Election Misinformation*, QUARTZ (Nov. 2, 2020), https://qz.com/1925272/the-risk-of-putting-warning-labels-on-election-misinformation/.

80. Rivero, *supra* note 79.

81. Pennycook et al., *supra* note 67, at 4945.

82. Rivero, *supra* note 79 (stating that while some studies do indicate that the tags work on that specific flagged post, "there's disagreement over whether [tags] can change the overall rate of misinformation spread on a platform"); *see also* Schwartz, *supra* note 9 (explaining that Trump supporters and young adults were more likely to be affected by the implied truth effect).

83. *Supra* note 82.

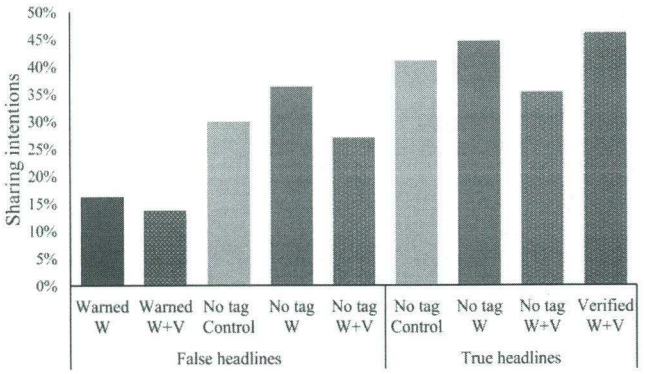
84. Claire Leibowicz & Emily Saltz, *Labeling Misinformation Isn't Enough. Here's What Platforms Need to Do Next.* PARTNERSHIP ON AI (Mar. 11, 2021), https://partnershiponai.org/labeling-misinformation-isnt-enough/.

85. Rivero, *supra* note 79.

86. *Id.*

implied truth effect because users understood that a post without a verified tag had not yet been examined by fact checkers.[87] The graphic below shows how the presence of verified labels and misinformation warnings on a post can influence a user's willingness to share the information:

Figure 4. (Color online) Fraction of Headlines That Participants in Study 2 Indicated They Would Consider Sharing, by the Tag Attached to the Headline (Top Row of *x*-Axis Label), the Experimental Condition (Middle Row of *x*-Axis Label), and the Headline's Veracity (Bottom Row of *x*-Axis Label)



Note. *W*, warning; *W+V*, warning + verification.

Figure 2. Willingness to Share.[88]

Misinformation tags contribute to both the implied truth effect and the backfire effect.[89] Yet social media platforms principally rely on these tags to inform users about misinformation in a post.[90] Without a change to the current misinformation policies, users on social media platforms will only continue to experience the implied truth and backfire effects.[91] Thus, the benefit that misinformation tags create by slightly increasing accurate perception of posts is outweighed by the

87. Pennycook et al., *supra* note 67, at 4945. A misinformation tag implies that the tagged post and other untagged posts have been evaluated whereas the verified post clarifies exactly which posts have been examined. *Id.*

88. Gordon Pennycook et al., *Figure 4* (illustration), *in* Pennycook, *supra* note 67, at 4952. This graphic illustrates that individuals are slightly less likely to share information that lacks a verified label (when verified labels are used) as compared to individuals who only encounter posts with misinformation labels. *Id.* Users were also slightly more likely to share verified posts than posts without a misinformation label when only misinformation labels were used and more likely to share verified posts than unverified posts. *Id.* The graphic also indicates that misinformation tags do help users decide whether to share a post because unmarked and verified posts were shared more often than posts marked as misinformation.

89. *See generally* Pennycook et al., *supra* note 67, at 4956 (describing the implied truth effect); Lewandowsky et al., *supra* note 11, at 119–20 (describing the backfire effect).

90. *See generally* Mukul, *supra* note 13; Gilbert, *supra* note 6; Hatmaker, *supra* note 22.

91. *See generally* Pennycook et al., *supra* note 67, at 4956; Lewandowsky et al., *supra* note 11, at 119–20.

detrimental effects of these phenomena.[92] These structural flaws in social media policies have led to political institutions criticizing both social media platforms and their "legal shield" against liability, Section 230.[93] With one political party claiming censorship and another claiming platforms are not doing enough to combat misinformation, social media platforms' immunities under Section 230 are no longer as secure as they once were.[94]

IV. PROTECTIONS OF AND RESPONSES TO SOCIAL MEDIA MISINFORMATION POLICIES

Section 230 comes from the Communications Decency Act of 1996.[95] This provision is considered to be the "internet's Magna Carta" because it facilitated the growth of the Internet.[96] The aims of the Communications Decency Act of 1996 were "promoting the development and free marketplace of the Internet, encouraging greater user control, and supporting voluntary use of filtering by providers."[97] Section 230 was inspired by a 1990s New York Supreme Court case, *Stratton Oakmont, Inc. v. Prodigy Services Co.*[98] In the 1990s, a brokerage firm sued Prodigy Services for defamation.[99] Prodigy Services was an internet services provider, and an anonymous user posted defamatory content on the site's message board.[100] The New York Supreme Court held Prodigy liable because the platform moderated users' posts and created content guidelines.[101]

Congressmen Ron Wyden and Christopher Cox worried that this court decision would disincentivize websites from moderating obscene content on their

92. Rivero, *supra* note 79.

93. Daisuke Wakabayashi, *Legal Shield for Social Media Is Targeted by Lawmakers*, N.Y. TIMES (Dec. 15, 2020), https://www.nytimes.com/2020/05/28/business/section-230-internet-speech.html.

94. *See generally* Anand et al., *supra* note 8.

95. Michelle P. Scott, *Section 230*, INVESTOPEDIA (Oct. 27, 2021), https://www.investopedia.com/section-230-definition-5207317. The goal of the Communications Decency Act was to "promote the development of the internet." *Id.*

96. Anand et al., *supra* note 8. Section 230 has been credited with allowing the Internet to grow and with protecting online expression because of the protections it grants to Internet platforms and users of those platforms. *Id.*

97. Robert C. Montañez, *Executive Order No. 13925: An Attempted Stop Sign on Our Global Cyber-Freeway*, GGU L. REV. BLOG (Oct. 2, 2020), https://ggulawreview.com/2020/10/02/executive-order-no-13925-an-attempted-stop-sign-on-our-global-cyber-freeway/.

98. Wakabayashi, *supra* note 93; *see also* Notice of Motion for Renewal and/or Reargument, Stratton Oakmont, Inc. v. Prodigy Servs Co., Index No. 031063/94, 1995 WL 17141954 (N.Y. Sup. Ct. July 5, 1995).

99. Wakabayashi, *supra* note 93.

100. *Id.*

101. *Id.*

platforms.[102] Wyden and Cox created Section 230 in part to immunize internet platforms from liability resulting from moderation of their own sites.[103] Wyden saw Section 230 as providing a "sword and a shield" for internet companies: a shield to block liability for user content and a sword to strike back against offensive content.[104]

Under Section 230's "Good Samaritan" provision, internet platforms cannot be held liable for content posted by users.[105] Nor can these platforms be liable for any good faith moderation of certain categories of content.[106] The provision does have some exceptions,[107] but for the most part Section 230 is broad enough to protect internet platforms from liability and to allow the platforms to offer a variety of different services.[108]

Since the passage of Section 230, courts "have repeatedly sided with internet companies."[109] Courts interpreted Section 230 as creating a broad immunity for internet platforms and held that it preempts other laws that impose liability on providers for third-party posts.[110] Section 230 acts as an affirmative defense.[111] And federal courts have "largely held that knowledge of unlawful content, editorial control, and decisions to remove or repost materials" do not revoke Section 230

---

102. *Id.* Given that Prodigy was held liable for moderation and content guidelines, other platforms would be cautious when moderating content on their own platforms because of the potential liability.

103. Casey Newton, *Everything You Need to Know About Section 230: The Most Important Law for Online Speech*, THE VERGE (Dec. 29, 2020), https://www.theverge.com/21273768/section-230-explained-internet-speech-law-definition-guide-free-moderation (explaining that Wyden and Cox wanted website owners to be able to moderate "without worrying about legal liability").

104. Wakabayashi, *supra* note 93.

105. 47 U.S.C.A. § 230(c)(1) (West). This provision states that no Internet platform will "be treated as the publisher or speaker of any information provided by another information content provider." *Id.*

106. *Id.* § 230(c)(2)(a). This provision states that that no Internet platform can be civilly liable for actions "voluntarily taken in good faith to restrict access to or availability of material that the provider . . . considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected." *Id.*

107. Bambauer, *supra* note 10. The provision provides exceptions for "violations of federal criminal law, wiretapping statutes, intellectual property rules, and . . . online sex trafficking." *Id.*

108. *Id.*

109. Wakabayashi, *supra* note 93.

110. Scott, *supra* note 95. Section 230 protects Internet platforms from "criminal sanctions, civil judgments, and injunctions." *Id.*

111. ELIZABETH BANKER, INTERNET ASS'N, A REVIEW OF SECTION 230'S MEANING & APPLICATION ON MORE THAN 500 CASES 5 (2020) https://internetassociation.org/wp-content/uploads/2020/07/IA_Review-Of-Section-230.pdf. While courts do often side with the internet companies, a true Section 230 judgment is relatively rare. *See id.* at 9–10. A significant number of cases (28%) are never decided on the merits of the Section 230 claim because courts dismiss them for other defects. *Id.* at 2. In only 42% of cases did a court make Section 230 the basis for its judgment. *Id.*

immunity from internet platforms.[112] *Zeran v. America Online, Inc.*[113] (AOL), considered to be one of the most important Section 230 rulings, established this editorial and moderation immunity.[114] In this case, an anonymous user advertised offensive shirts on an AOL bulletin board and provided the plaintiff's phone number as a contact.[115] An anonymous user continued to create similar posts, and the plaintiff eventually sued AOL.[116] The Fourth Circuit held that Section 230 prevents an internet service provider from being liable for "its exercise of a publisher's traditional editorial functions – such as deciding whether to publish, withdraw, postpone or alter content."[117] Thus, Section 230 protects the moderation of content on a platform.[118]

Section 230 governs social media because these platforms are considered to be "interactive computer services" under the provision.[119] Because Section 230 applies to social media platforms,[120] the provision protects these platforms from liability for user posts on their platforms *and* the removal of those posts.[121] In other words, Section 230 allows these sites to label posts or ban users who do not follow the platform's misinformation policies.[122] But this provision—and the immunities it grants to social media platforms to moderate content—has come under attack by Congress, Presidents, and numerous state governments.[123]

A. Congressional Responses

Both political parties have attempted to reform Section 230.[124] While Republicans and Democrats in Congress agree that reform is necessary, they

112. Montañez, *supra* note 97.

113. Zeran v. Am. Online, Inc., 129 F.3d 327 (4th Cir. 1997).

114. Eric Goldman, *The Ten Most Important Section 230 Rulings*, 20 TUL. J. TECH. & INTELL. PROP. 1, 3 (2017).

115. *Zeran*, 129 F.3d at 329.

116. *Id.* AOL did remove the posts as the plaintiff informed the platform, but the anonymous user continued to create more posts to replace the deleted posts. *Id.*

117. *Id.* at 330.

118. *Id.* at 331.

119. Bryan Pietsch, Isobel Asher Hamilton & Katie Canales, *The Facebook Whistleblower Told Congress It Should Amend Section 230, the Internet Law Hated by Both Biden and Trump. Here's How the Law Works,* INSIDER (Oct. 6, 2021, 10:39 AM), https://www.businessinsider.com/what-is-section-230-internet-law-communications-decency-act-explained-2020-5.

120. *Id.*

121. *See* Wakabayashi, *supra* note 93.

122. Pietsch et al., *supra* note 119.

123. *See generally id.* (describing why political parties want to reform or repeal Section 230).

124. *See generally* Anand, *supra* note 8 (exploring the arguments for reform from both Republicans and Democrats and explaining the substance of the bills that have been proposed); *see also* Pietsch et al., *supra* note 119.

disagree on the form these reforms should take.[125] Republicans largely believe that social media platforms are "censoring conservative viewpoints."[126] Because Section 230's policy is to foster diverse discourse,[127] Republicans argue that misinformation policies targeting conservative speech refute the provision's policy.[128] Republican reformation attempts focus on the policies of these platforms and aim to punish discriminatory content moderation.[129] Conversely, Democrats want Section 230 amendments to impose liability on companies that do not do enough to moderate harmful content.[130] Democrats argue that the current policies of social media platforms are ineffective at tackling the misinformation problem.[131] Put differently, Democratic proposals are attempts to force platforms to remove more content.[132] To address their concerns, both political parties have introduced bills to amend Section 230.[133]

Regardless of political affiliation, proposals for reform generally fall into one of two categories: a "carveout" approach and a "bargaining chip system."[134] The carveout approach advocates creating immunity exceptions for certain types of content.[135] An example of this type of approach is the Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) that was passed in 2018.[136] Under the Act, websites do not have Section 230 immunity for sex trafficking charges or conduct that promotes prostitution.[137] The bargaining chip approach proposes only immunizing platforms from liability once they meet certain standards.[138] An example of this type of approach is a bill proposal that would give internet companies immunity if the company submitted audits showing their policies are politically neutral.[139]

---

125. Wakabayashi, *supra* note 93.

126. Spangler, *supra* note 43.

127. 47 U.S.C.A. § 230(b) (West).

128. Wakabayashi, *supra* note 93.

129. Newton, *supra* note 103.

130. Pietsch et al., *supra* note 119.

131. Wakabayashi, *supra* note 93. The policies are considered ineffective since misinformation continues to exist and spread on social media platforms in large quantities. *Id.*

132. Newton, *supra* note 103.

133. Anand, *supra* note 8. This website summarizes bills lawmakers proposed to repeal or amend Section 230. It also states the bill name, sponsors, date introduced, and status. At the time this article was written, the list had been last updated on December 6, 2021.

134. Newton, *supra* note 103.

135. *Id.*

136. *Id.*

137. *Id.*

138. *Id.*

139. Wakabayashi, *supra* note 93.

B. Executive Responses

Presidents have also criticized Section 230.[140] In fact, during their 2020 presidential campaigns, both President Donald J. Trump and President Joseph R. Biden called for Section 230 repeals.[141] After Twitter placed misinformation tags on President Trump's tweets in May 2020, he issued an executive order that targeted Section 230 protections for social media platforms.[142] Executive Order on Preventing Online Censorship (Executive Order No. 13925) purported to not immunize online platforms that "engage in deceptive or pretextual actions . . . to stifle viewpoints with which they disagree."[143] President Biden revoked the executive order when he took office.[144] The revoked executive order ultimately did not affect Section 230, but it did incentivize lawmakers to look at the provision more closely and to start targeting its protections.[145]

While President Biden has not issued any executive orders affecting Section 230, he has accused social media platforms of "'killing people' by facilitating the spread of misinformation about coronavirus vaccines."[146] In January 2020, President Biden also proposed repealing Section 230.[147] But he has not created a political agenda addressing Section 230 since being elected.[148]

C. State Responses

While these state bills are "doomed to fail while . . . Section 230 . . . is in place," many state lawmakers have also pushed for reforms for social media protections.[149]

---

140. Scott, *supra* note 95.

141. Anand, *supra* note 8.

142. Pietsch et al., *supra* note 119.

143. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (May 28, 2020), *revoked by* Exec. Order No. 14,029, 85 Fed. Reg. 27,025 (May 14, 2021). This executive order applied to social media platforms since they were considered content publishers. *Id.* The executive order purported to remove Section 230 protections from social media platforms in the applicable circumstances. *Id.*

144. J. Fingas, *President Biden Revokes Trump Order Limiting Social Media Protections*, ENGADGET (May 15, 2021), https://www.engadget.com/biden-revokes-trump-social-media-executive-order-170713081.html. Biden did not explain why he revoked the executive order. *Id.* Prior to its revocation, the executive order was never "meaningfully enforced" because the Trump Administration was sued by activists over claims that the executive order suppressed free speech. *Id.*

145. Pietsch et al., *supra* note 119.

146. Spangler, *supra* note 43.

147. Newton, *supra* note 103.

148. *Id.* While Biden has not proposed any specific agenda addressing Section 230, one of his advisors did suggest repealing Section 230 and replacing it with a new law. *Id.*

149. Anthony Izaguirre, *GOP Pushes Bills to Allow Social Media 'Censorship' Lawsuits*, AP NEWS (Mar. 7, 2021), https://apnews.com/article/donald-trump-legislature-media-lawsuits-social-media-848c0189ff498377fbfde3f6f5678397. Experts argue that Section 230 makes these bills

These state bills are largely proposed by Republican state lawmakers, and many of the proposals focus on creating civil suit provisions for censorship on internet platforms.[150] Republican legislators introduced legislation in approximately two dozen states.[151] The Heartland Institute estimated in September 2021 that "70 bills in 30 states are challenging 'big tech' censorship."[152]

Texas and Florida both passed laws targeting content restrictions on Internet platforms.[153] The Florida law, which was stayed in June 2021 by a federal judge, penalized platforms that restricted politicians' posts.[154] Similarly, the newly enacted Texas law grants citizens the right to sue an Internet platform if that platform bans the individual because of his or her political viewpoint.[155] Facebook, Twitter, and YouTube are suing Texas to stay this law.[156]

D. Public Responses

Researchers have not discovered widespread evidence of conservative censorship on internet platforms.[157] Despite that finding, one survey indicates that "roughly three-quarters of U.S. adults say it is very (37%) or somewhat (36%) likely that social media sites intentionally censor political viewpoints."[158] While majorities from both political parties believe censorship may occur, the belief is most common among Republicans.[159] In 2018, 85% of Republicans and Republican-leaning independents believed social media websites censored political viewpoints; in 2020, 90% believed it was somewhat likely censorship occurred.[160] A similar group of studies involving 34 participants found that many participants believed Internet platforms should not moderate user content, and approximately

---

unconstitutional. *Id.* However, lawmakers continue to press these state bills. *Id.* Most of the bills are being pushed by Republican lawmakers who believe that the Section 230 liability shield should be removed or updated because of current censorship claims. *Id.*

150. *Id.*

151. *Id.*

152. Jessica Guynn, *Facebook, YouTube and Twitter Strike Back, Sue over Texas Social Media Censorship Law*, USA TODAY (Sept. 22, 2021, 1:34 PM), https://www.usatoday.com/story/tech/2021/09/22/facebook-youtube-twitter-lawsuit-texas-conservatives-censorship-law/5803509001/.

153. *Id.*

154. *Id.*

155. *Id.*

156. *Id.*

157. Izaguirre, *supra* note 149.

158. Emily A. Vogels et al., *Most Americans Think Social Media Sites Censor Political Viewpoints*, PEW RSCH. CTR. (Aug. 19, 2020), https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/.

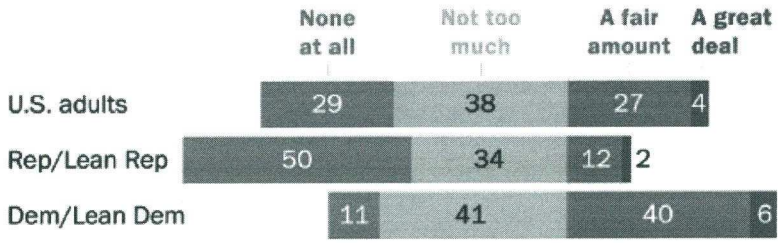
159. *Id.*

160. *Id.*

half reacted to labels "with skepticism or outrage."[161] The participants with those beliefs were Republicans or Independents.[162]

The public as a whole is divided on whether social media platforms should fact check posts at all.[163] Approximately 66% of Americans "have not too much or no confidence at all in social media companies being able to determine which posts . . . should be labeled as inaccurate or misleading."[164] The graphic below highlights the breakdowns of the survey in this area:

% of U.S. adults who say they have ___ (of) confidence in social media companies to determine which posts on their platforms should be labeled as innaccurate or misleading



Note: Strongly/somewhat approve or disapprove responses are combined. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted June 16-22, 2020.
"Most Americans Think Social Media Sites Censor Political Viewpoints"

Figure 3.[165]

Thus, as social media platforms continue to try to halt the spread of misinformation, such policies may instead instill a lack of distrust in at least a portion of the American public.[166] In fact, these policies have led in part to the creation of new social media platforms.[167] Sites like Parler became more popular

161. Leibowicz & Saltz, *supra* note 84.

162. *Id.*

163. Vogels et al., *supra* note 158.

164. *Id.*

165. Pew Rsch. Ctr., *Wide Partisan Differences in Views About Social Media Companies Labeling Posts from Elected Officials As Misleading Or Inaccurate* (illustration), *in* Vogels et al., *supra* note 158. This graphic demonstrates that people from both political parties have either none or little faith in the ability of social media companies to accurately label misinformation. *See id.*

166. Vogels et al., *supra* note 158.

167. Queenie Wong & Andrew Morse, *Parler Returns Online After Monthlong Absence: Here's What You Need to Know*, CNET (Feb. 16, 2021, 5:22 PM), https://www.cnet.com/news/parler-returns-

among Republicans "amid allegations that Twitter, Facebook, and other social networks harbor anti-conservative bias."[168] Additionally, President Trump created his own social media platform, Truth Social, which launched in February 2022.[169] These social media sites seem to have less restrictive content guidelines.[170] For example, Parler neither uses independent fact checkers nor tags posts.[171] This growing distrust of traditional social media platforms, coupled with legislative and presidential attacks against Section 230, indicate that social media platforms may no longer enjoy—and perhaps should not enjoy—the same immunity in the coming years.

V. THE UNCERTAIN FUTURE OF SOCIAL MEDIA IMMUNITY

No easy answer exists to the problem of misinformation. This misinformation problem is only exacerbated by current social media policies.[172] Social media platforms principally rely on algorithms to address the sheer amount of misinformation on their platforms, yet these same algorithms also feed users misinformation.[173] Algorithms are also not entirely reliable and have misflagged credible posts as misinformation.[174] Further, misinformation flags are one of the main tools by which social media platforms inform users about misinformation.[175] These flags may not only result in misinformation reaching a broader audience[176] but can also strengthen a user's belief in misinformation.[177] The implementation

online-after-month-long-absence-heres-what-you-need-to-know/ (describing the social media site Parler).

168. *Id.*

169. Bobby Allyn, *Truth Social, Donald Trump's Social Media App, Launches Year After Twitter Ban,* NPR (Feb. 22, 2022, 5:14 AM), https://www.npr.org/2022/02/22/1082250086/truth-social-donald-trumps-social-media-app-launches-year-after-twitter-ban.

170. Wong & Morse, *supra* note 167. Parler does not have guidelines prohibiting hate speech, but it does have some content guidelines. *Id.* For example, the site does not allow "terrorists, spam, unsolicited ads, pornography, threats to harm, blackmail, and content that glorifies violence against animals." *Id.* The guidelines also prohibit obscene content, which is defined as sexual, offensive content. *Id.*

171. *Id.*

172. *See* Ortutay & Seitz, *supra* note 12.

173. *See generally* Spangler, *supra* note 43; Bessi & Quattrociocchi, *supra* note 2, at 36; Meserole, *supra* note 32.

174. Meserole, *supra* note 32.

175. *See generally* Mukul, *supra* note 13; Gilbert, *supra* note 6; Hatmaker, *supra* note 22.

176. Devitt, *supra* note 53 (explaining that some flagged tweets have reached a broader audience than unflagged tweets).

177. *See generally* Pennycook et al., *supra* note 67, at 4956 (describing the implied truth effect); Lewandowsky et al., *supra* note 11, at 119–20 (describing the backfire effect).

and ineffectiveness of these policies have also negatively impacted the public's trust of social media platforms.[178]

Lawmakers shift the blame for the misinformation problem on the immunities that Section 230 grants to social media platforms.[179] Even though research shows that current social media policies are ineffective and counterproductive,[180] Section 230 still shields platforms from liability for ineffective content moderation.[181] Faced with this prospect of social media immunity, lawmakers have taken steps to weaken or repeal Section 230.[182] Even presidents have advanced their intentions to attack this provision.[183] Consequently, social media platforms and their policies now stand on shaky ground. While no changes have yet been made, as the misinformation problem continues to run rampant, the landscape of immunity once granted to the internet may soon change dramatically.

178. *See generally* Vogels et al., *supra* note 158 (explaining that Republicans largely view social media platforms as censoring them and that the public seems to mistrust the ability of social media platforms to moderate misinformation).

179. *See generally* Bambauer, *supra* note 10; Anand, *supra* note 8.

180. *See generally* Rivero, *supra* note 79; Lewandowsky et al., *supra* note 11, at 119.

181. *See generally* Pietsch, *supra* note 119.

182. Newton, *supra* note 103 (describing congressional attempts to reform or repeal Section 230); *see also* Anand, *supra* note 8 (summarizing many of the proposed bills).

183. Anand, *supra* note 8 (explaining both President Trump's and President Biden's stances on Section 230).