Summer 6-29-2021

Combatting Misinformation: Effective Twitter Responses to The Removal of Civil Liability Protection in Section 230

Phillip Severson

Combatting Misinformation: Effective Twitter Responses to The Removal of Civil Liability Protection In Section 230

Phillip Severson
EMPA 396 Graduate Research Project in Public Management
Golden Gate University
San Francisco, California
Dr. Mick McGee & Dr. Joaquin Gonzalez, III
June 29, 2021

TABLE OF CONTENTS

ABSTRACT

Growing concern surrounding the harmful effects of widespread misinformation on social media platforms, policy-makers and influencers, including the Department of Justice, have proposed reforms to Section 230 such as the removal of civil liability protection to incentivize more effective business practices at preventing widespread misinformation on social media platforms. This study gained insight into the perceived effectiveness of business practices Twitter may undertake to significantly reduce volumes of misinformation incentivized by the removal of civil liability protection from Section 230. In particular, a Theory of Change was examined in which Twitter Inc. would attempt to significantly reduce misinformation through an increase content moderation; improvement of user verification and authentication processes; and intensification of public transparency practices. Relevant literature indicated that increased moderation could have harmful effects upon the user experience, such as over-moderation, but showed promising benefits in the application of improved user verification and public transparency practices. Quantitative insights from survey respondents generally validated each of business practices encapsulated in the assumptions from the Theory of Change, but qualitative insights from the survey respondents and interviews with selected subject matter experts identified significant challenges, limitations, and caveats regarding the implementation of each assumption. Of each assumption, the act of improving user verification and authentication process was expected to provide the greatest benefit. If civil liability protection is removed from Section 230 in efforts to incentivize more effective business practices to reduced and limit misinformation, it is recommended that Twitter increase content moderation, improve user verification, and intensify its public transparency practices, however with several caveats informed by qualitative perspectives.

CHAPTER 1

BACKGROUND OF THE PROBLEM

In recent years, the term "fake news" was often attributed to former U.S. President Trump in his descriptions of several mainstream media news outlets purported to have provided intentionally inaccurate news stories due to political leanings. While few could argue that the idea that fake news is wrong, perhaps the idea of fake news from mainstream media outlets as misinformation is far too narrow. Misinformation is generally defined among several studies as inaccurate information presented in the form of truth with an ability to cause public harm (The Information Society Project, Yale Law School, 2017, p. 5).

There are several causes to the recent explosion of misinformation. Lee (2021) states, "The erosion of public trust in traditional news sources creates a vacuum filled by misinformation" (p. 85). As a result, opportunities lie within platforms such as Facebook, Reddit, and Twitter, which provide not only connections among people, but the ability to share news and information among users. Butler (2018) notes that the combination of post-modern thinking—in which subjective ideas outweigh objectivity—and the rate at which Americans get their news through social media is recipe for a misinformation foothold within culture (Butler, 2018, p. 427).

Several studies show the devastatingly harmful effects of consuming large amounts of misinformation among the public. A peer-reviewed Harvard study by Ognyonova et al. (2020) shows that exposure to misinformation has significant impacts upon public trust in certain institutions such as mainstream media and the government (p. 3). Other studies show that misinformation can have devastating effects upon public health as with misleading information surrounding the COVID-19 epidemic (Gupta et al., 2020, p. 2). Further, other studies show how

misinformation can lead to swaying of public opinion towards polarization and public discord (The Information Society Project, Yale Law School, 2017, p. 5). This has become such a global issue that other nations like Russia, China, and France have attempted to address widespread misinformation in their own respects according to Levush (2020) at the Library of Congress.

In light of the discussion above, U.S. Federal government and policy makers have entertained legislative and policy actions in attempting to mitigate and/or lessen the proliferation of misinformation, primarily focusing efforts on 47 U.S. Code § 230. As it stands, 47 U.S. Code § 230 ("Section 230") of the Communications Decency Act (CDA) of 1996 is credited for the boom in growth of internet-based industries like social media companies (Armjio, 2021, p. 3). In brief, Section 230 promotes a U.S. public policy aimed at development of the internet and "interactive computer services" and media, and it provides a "Good Samaritan" civil liability protection for content moderation practices for user-generated media proliferated on their platforms. ISPs maintain a level of civil liability protection for their efforts in moderating user-generated content that is "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected" (47 U.S. Code § 230(c)(2)(A)). As such, Section 230 is attributed to the rise of misinformation through ISPs (Butler, 2018, p. 435).

Some existing conversations surrounding legislative and policy actions that attempt to address misinformation through Section 230 will categorize misinformation as a form of media that ISPs may legally moderate according to the definitions outlined in 47 U.S. Code § 230(c)(2)(A). As such, social media companies have business practices that moderate content of various objections, and companies like Facebook and Google have attempted to mitigate effects of misinformation head on (Torres et al., 2018). Social media companies employ a variety of

techniques to moderate misinformation, but some critics may argue that these measures are not effective enough to address the extensive, global volume of misinformation emanating (among other issues) from social media platforms (Armijo, 2021, p. 3).

Legislative and policy efforts aimed at reducing and mitigating the proliferation of misinformation through ISPs will often cite, reference, or target 47 U.S. Code § 230 and investigate its implications towards widespread misinformation. For example, a study performed by the Office of the Attorney General of the Department of Justice (2020) resulted in recommendations of what it believes to be multiple courses of action that policy makers could implement to combat misinformation proliferated through ISPs. Among these is the potential removal of civil liability protections afforded to ISPs by 47 U.S. Code § 230.

Specifically, the Office of the Attorney General of the Department of Justice (2020) recommended that the removal of protections from civil lawsuits, in which ISPs are protected by 47 USC §230(c)(2), would allow for "civil enforcement actions brought by the federal government." Currently, 47 USC §230(c)(2) under the "Protection for 'Good Samaritan' blocking and screening of offensive material" clause reads:

> "No provider or user of an interactive computer service shall be held liable on account of—(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or (B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1)."

As such, this study looks to identify the potential courses of action Twitter may undertake as a result of the said removal of civil liability protection from Section 230.

STATEMENT OF THE PROBLEM

Legislation towards for the removal of civil liability protection clause 47 U.S. Code § 230(c)(2) of Section 230 has potentially significant implications on the current volumes of misinformation proliferated from ISPs, in particular to Twitter Inc. These implications include potential significant reduction in proliferation of false and misleading content on Twitter and changes to Twitter's business practices, such as increased content moderation, user verification processes, and public transparency.

PURPOSE OF THE STUDY

The purpose of this study is to examine the perceived efficacy of business changes possibly to be undertaken by Twitter Inc. to combat misinformation proliferating from its platform resulting from a policy change in which civil liability protection (as defined under 47 U.S. Code § 230(c)(2)) is removed from Section 230. In particular, this study examines Twitter's current content moderation practices against misinformation, and it seeks to identify the perceived efficacy of the company's potential courses of action in policy changes to users' identity verification processes, user-generated content moderation, and Twitter transparency practices in relation to the volumes of misinformation proliferating from Twitter.

For clarification, this study does not advocate for the removal of civil liability protection from 47 U.S. Code § 230. Further, this study does not adopt a partisan stance associated with political controversies surrounding Section 230, nor seek to limit its scope (Cheah, 2020, p. 194). Further still, while issues surrounding misinformation and biased selective censoring by social media companies may overlap, this study does not advocate for any proposed bill, such as a bill by Senator Josh Hawley in June 2020 titled "Limiting Section 230 Immunity to Good Samaritans

Act" which targets 47 U.S. Code § 230(c)(2) and is focused on social media and political neutrality issues (Armijo, 2021, p. 5).

SIGNIFICANCE OF THE STUDY

Combatting misinformation has gained bipartisan agreement throughout the U.S. political spectrum (Klein, 2020, p. 59). It is recognized that the mass ingestion of misinformation by the population has significant effects, for example, by creating a population of an "uninformed electorate which casts votes based on incomplete, biased, or fraudulent fact reporting" in addition to other ramifications (p. 47). Sharing of misinformation may erode "public trust in institution, and on social harmony" (Levush, 2019). Most readily apparent, misinformation undermines the function of the press (Butler, 2018, p. 426).

It is also the policy of the US government to "preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal and State regulation" according to 47 USC §230(b). Therefore, discussion towards potential regulation upon any form of media (except for decency standards of explicitly illicit content) has major implications for ISP business practices and internet-based innovation. Therefore, policy makers at the federal level are the first and foremost benefiters of this information. However, the greatest contributions will be to Social Media ISPs (i.e. Twitter, Facebook, Reddit, Telegram, etc.), users of the platforms, and potentially private mainstream media (CNN, MSNBC, Fox News, etc.).

Misinformation is important to public health as well. Studies showed that even medical professionals resort to social media most for obtaining information regarding COVID-19 above all other mediums (Gupta, et al. p. 7). However, Gupta, et al. (2020) notes that compared to other types of media, social media was the largest perceived proliferator of misinformation

regarding COVID-19 as identified by surveyed medical professionals (p. 2). The implications of sharing widespread misinformation regarding health concerns has obvious negative implications for the general public.

No doubt, due to the global reach of major social media companies, these policies will affect information flow into other countries, for example France's existing Twitter misinformation policy regarding voting issues (Twitter Help Center, 2021).

RESEARCH QUESTION AND SUBQUESTIONS:

This study aims to provide insight on the perceived causal relationship between the removal of civil liability protection defined in 47 Section 230(c)(2) and its impacts on the volume of proliferation of misinformation through Twitter Inc. should the company undertake specific actions outlined in this research. This study will also attempt to determine perceptions as to whether Twitter is currently providing enough and/or effective content moderation towards mitigating the proliferation of misinformation from its platform and gain insight into the perceived impact of the removal of civil liability protection defined in 47 Section 230(c)(2) upon the users, user-generated content, and Twitter Inc.

THEORY OF CHANGE AND ASSUMPTIONS

The Theory of Change (TOC) and assumptions for this study are as follows: If Internet Service Providers (ISPs) civil liability protection is removed (as defined by 47 USC §230(c)(2)) from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation; then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation; and then an intensification of public transparency by Twitter will significantly reduce misinformation. The assumptions behind this theory are as follows:

Assumption 1 (A1): If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation.

Assumption 2 (A2): If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation.

Assumption 3 (A3): If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an intensification of public transparency by Twitter will significantly reduce misinformation.

LIMITATIONS

Since this legislation has only been proposed and/or recommended, the actual impacts of the study can only gain insight into potential outcomes of Twitter's response as well as scholarly and user perceptions of those actions. While this study alludes to impacts of Internet Service Providers (ISPs) as a whole (with special focus on social media companies) there are significant inherent differences in the companies regarding issues such as their audience, consumer appeal, business practices, economic, and political impacts to U.S. society and globally. As such, this study may allude to situations and insights gained from other ISPs which may help qualify issues affecting Twitter Inc.

Further, this study intentionally does not provide in-depth insight into implications regarding potential First Amendment rights issues nor the application of similar concepts upon

privately-owned platforms, which may or may not be affected by changes to Section 230. Nor does this provide in-depth insight into the legal viability of civil liability and other legal ramifications should ISPs like Twitter become liable for user-generated content on its platform. This study does not provide in-depth insight into other recommended supplementary and/or alternative actions that could potentially benefit the overall goal of reducing misinformation on the Twitter platform, for example, the recommendation by the Office of the Attorney General of the Department of Justice (2020) to clarify ambiguous terminology in 47 USC §230(c)(2)(A) such as "otherwise objectionable."

Most importantly, this study does not intend to provide in-depth insight into issues regarding controversial content moderation such as whether various ISPs, including Twitter, moderate political opinions and biases. However, this study acknowledges that facts may be presented in ways that opposing viewpoints may potentially and inaccurately label as misinformation, alluding to Torres et al. (2018) statement that there is a "shifting definition of fake news, changing the focus from satirical commentary to willful malevolence" (p. 3983).

DEFINITION OF TERMS

This study is based on terminology derived from 47 USC §230, which includes additional definitions concerning the internet and ISPs. For example, internet social media companies fall under the umbrella of ISPs, however not all ISPs are of the social media sort.

Users may publicly share content, or "proliferate information" in various ways. Twitter specifically allows users to share publicly viewable messages, known as "tweets." Tweets may be comprised of words, links, embedded articles, embedded videos, images, polls, hashtags, etc. Tweets may also reference other users' tweets and continue to share that content. Twitter users may also "follow" other users, which allows them to subscribe to notifications of activities

performed by the "followed" Twitter user. Therefore, publicly-posted activity conducted by a user who is followed by many will likely generate notifications and content of that activity to each of its many followers. For the sake of this study, a user who tweets information also "proliferates" information. Therefore, it is important to note that proliferation of misinformation refers to user-generated tweets, which are not originated from Twitter Inc. nor its employees.

EXPECTED IMPACT OF THE RESEARCH

Public perception is a powerful tool, and insights gained from this study has potential value in guiding and assisting the development of various public, private, and internal policies towards moderation of misinformation. As a proof of concept example, Twitter plans to relaunch its verification policy in 2021 as a result of 22,000 survey responses regarding its Blue Badge (Twitter Inc., 2020).

Going forward, these policies may impact other user and information verification practices, such as authentication protocols, algorithms, terms of service, and legal perspectives. This may also shape how Twitter will continue its public messaging such as public announcements surrounding content moderation, including misinformation. Similarly, other ISPs of like function may benefit in like manners towards refining industry practice.

CHAPTER 2: LITERATURE REVIEW

INTRODUCTION

The following literature review introduces key points from scholars and academics surrounding elements in this study's Theory of Change. In application to Twitter, the literature review discusses the volume problem for the ISP industry and current moderation practices employed. It also discusses implications of enforcing stricter moderation practices and potential impacts upon users. Finally, it discusses business transparency practices.

THE VOLUME PROBLEM AS THE BACKDROP

The major problem that social media companies encounter regarding content moderation for misinformation is simply, yet complexly, the size of data with which they must tackle. Cramer (2020) states that there is so much content, that no matter the current business practice employed (algorithms, policies, artificial intelligence, user reporting, etc.), social media companies are unable to keep up with the demand for combatting this issue (p. 135).

Goodyear (2020) attributes the explosion of misinformation over the internet to the removal of a publisher in which traditional, publicized content was vetted (p. 281). Further amplified on social media, users usually only share articles which "support a position" (p. 281). As an example, this has become so prolific, Allcott, et al. (2019) notes that surrounding the 2016 election, the consumption of major news site articles was far outnumbered by the consumption of misinformation websites in as much as two-thirds (p. 2). Surrounding the election, a study was performed by Buzzfeed which showed that "fake news" received more interactions than information from generally respected news sources as the New York Times and The Washington Post (Goodyear, 2020, p. 280).

SOCIAL MEDIA MODERATION PRACTICES RESULTING FROM SECTION 230

As a result, social media companies have attempted to combat this in various ways, as permitted by Section 230 to moderate content. For example, Allcott, et al. (2019) highlights misinformation "algorithmic and policy changes" contrasting the world's top two social media companies, Twitter and Facebook (p. 1). Leading up to the 2016 election, the quantity of misinformation for the two ISPs increased at similar rates (Alloctt, et al, 2019, p. 4). Allcott, et al (2019) noted that Facebook's initial efforts at combatting misinformation shortly after the election resulted in a significant drop of consumption of misinformation on its platform (p. 2). It was initially assessed that the drop was apparently attributed to Facebook's policy changes that succeeded the election (p. 2). However, the consumption of misinformation on Twitter continued while Facebook's consumption decreased (p. 4). Allcott, et al (2019) asserts that the Twitter userbase is more political in nature, potentially contributing to the reason why misinformation continued to grow shortly after the 2016 election (appendix p. 4). This demonstrates that there isn't a one-size fits all business model for ISPs due to their functionality and customer base.

No prescreening. There has been confusion regarding the relationship between social media companies and content that is distributed within the platforms. Cramer (2020) acknowledges a blanket provision by Section 230, which does not require user-generated content to be prescreened (p. 125). One can analogize a newsstand which is "not generally held liable for the content they distribute" (a distributor) and has little "editorial control" (a publisher) over what is distributed (Goodman & Wittington, 2019, p. 2). However, this relationship does not clearly define the internet-based relationships because ISPs acted in similar respects as both distributors and publishers (p. 2). As a result, ISPs' policies may come into question as the

dynamic of moderation is not clearly understood and content reviewers are already behind the power curve.

Policies. Cramer (2020) acknowledges that social media companies will attempt to reduce the distribution of misinformation through changes in policies like how YouTube and Facebook attempted to reduce visibility of misinformation regarding vaccines (p. 134). Twitter also adopts this approach for tweets that meets certain criteria to Twitter's COVID-19 misinformation policy (Rojo, 2021). However, aside from the COVID-19 misinformation policy, there is no specific, general misinformation policy except for the "Reporting false information in France" policy and the "Authenticity" policy, under which general misinformation could reasonably apply (Twitter Help Center, 2021). However, Twitter discusses misinformation further under its "Twitter Moments Guidelines and Principles" in which its curation team identifies tweets that "aim to uphold high standards of accuracy, impartiality, and fairness" in curation and will therefore include accurate Tweets in its Moments conversation above less accurate tweets (Twitter Help Center, 2021).

As an example of misinformation moderation at work, according to the Camino Rojo (2021), who is the Head of Public Policy, Government & Philanthropy at Twitter, the company implemented the COVID-19 misleading information policy in March 2020, in which Twitter's team removed 8,493 tweets by the time of publishing. In stark contrast, Twitter's algorithms "challenged 11.5 million accounts which were targeting discussions around COVID-19 with spammy or manipulative behaviors" (Rojo, 2021).

Fact Checkers. Lim (2018) as cited in Allcott, et al. (2019) acknowledges the challenge of using fact checkers is their low "inter-rater reliability" in determining what is inauthentic (p. 6). This may be attributed to opposing worldviews and limited capability of being objective.

Cramer (2020) supports this by stating that complexities in language combined with massive amounts of data makes the task of scrutinizing "objectional" content very challenging, if not impossible (p. 135).

Flagging. Facebook used to combat misinformation by "flagging inaccurate stories as 'Disputed,'" however this was perceived to have only caused a modest reduction in the trust of the content by viewers (Clayton et al., 2019, as cited in Alcott, H. et al., 2019). Conversely, Alcott, et al. (2019) noted that the lack of a presence of a "false" tagging can cause a story to be interpreted as far truer than it actually may be (p. 1).

Similarly, Twitter Help Center (2021) states that Twitter may "apply a label and/or warning message" in a potentially misleading post, in addition to other actions. Such a post became famous with the tagging of former President Trump's Twitter profile regarding mail-in ballots (Chea, 2020, p. 193). In addition to this, Vertstraete et al. (2017) calls to light an additional "crowd-sourcing" architecture, a user-reporting mechanism, which works in concert with other methods to validate the information (p. 27). However, Cramer (2020) believes this technique is overly limited in its span as it attributes to "over-reporting of dubiously harmful content while missing many posts that really are objectionable" (p. 135).

IMPLICATIONS OF MORE CONTENT MODERATION

Moderation of misinformation is not clear, cut and dry. In fact, it can be quite confusing when considering what constitutes information that is false, or partially false. For example, participants in the Information Society Project at Yale Law School (2017) workshop stated that moderating explicitly inaccurate misinformation can be addressed by algorithms, but misinformation has aspects of truth that is "buried under speculation, hyperbole, defamation, and spin" and is much more challenging (p. 10). Adding to this, Vertstraete et al. (2017) states that

propaganda misinformation, which "mixes fact and fiction," often eludes common solutions regarding misinformation (p. 13). This may shed light as to why out of 11.5 million accounts that were challenged, only several thousand tweets were removed with regards to Twitter's efforts surrounding COVID-19 misinformation, as previously mentioned in the Twitter COVID-19 misinformation policy (Rojo, 2021). As such, Goodman and Wittington (2019) suggest that if ISPs are put in positions where they must moderate content based on political and potentially ambiguous ends, this could have negative results (p. 4). For example, ISPs might drop the idea of moderation altogether and act strictly as a distributor of content, therefore making their "platforms open to [more] misinformation, sexually explicit content, and harassment" (p. 4).

As it stands with moderation practices, Cheah (2020) states that the scale at which social media companies operate has significant complexities that "prescriptive content moderation" can't address without causing further issues (p. 216). For example, Armijo (2021) commented on Twitter's deployment of automation, which was intended to combat material proliferated by white supremacists but inadvertently resulted in disrupting tweets from political conservatives (p. 19). This supports the assessment that "stronger notice and take-down enforcement regime" would likely result in over moderation (The Information Society Project, Yale Law School, 2017, p. 9). Lee (2021) would agree, stating that if social media companies "could be liable for all user-generated content that they moderated, they wouldn't moderate anything at all" (p. 87).

USER VERIFICATION PRACTICES

Stanford's 2017 Practicum Research Team (2017) conducted a study on misinformation policies and noted that verification tools within social media platforms may be a critical component to combatting misinformation (p. 131). In particular, the study noted that Twitter's use of the "blue verification badge" helps users trust the content coming from the user's account

(p. 131). While this verifies the user, this does not verify the validity of the information shared (p. 131).

Nonetheless, a population of verified users has its benefits. For example, Torres, et al. (2018) conducted a study of among social media users and noted differentiating news verification behaviors among users of social media profiles. Torres et al. (2018) noted that authentic users who carefully construct an online persona towards a "desired public image" (and who are likely to share information) are "more likely to engage in information verification behaviors" for fear of reprisal, judgement, or negative feedback from its network of social media connections (p. 3981). Additionally, Torres et al. (2018) found that a "trust in network" had a correlating positive increase in news verification behaviors by users (p. 3983).

The 2017 Practicum Research Team (2017) from Stanford would agree, as their surveys showed that "users associate the blue verification symbol with truthfulness and trustworthiness (p. 120). It was further recommended that Twitter offer blue verified badge option to be permitted to all users (p. 111). However, at the time of this study in 2021, the program for the blue verification badge is on hold tentative a relaunch in 2021 (Twitter Inc., 2021). Twitter Inc. (2021) stated that the new verification policy "will lay the foundation for future improvements by defining what verification means, who is eligible for verification and why some accounts might lose verification to ensure the program is more equitable." Further, the process for applying for a verification badge includes correlating identity through links and other supporting materials (Twitter Inc., 2021).

BUSINESS TRANSPARENCY PRACTICES

As top social media companies share geographic concentrations and similar business models, Twitter often gets lumped in with the "Big Tech" label and is not immune from such

public perceptions. Cramer (2020) acknowledges that social media companies exhibit "inconsistent attitudes" regarding poor content on their platforms (p. 126). Because of inconsistent attitudes, it may be difficult for a social media company to be seen as "transparent." For example, several major news articles including USA Today note Facebook CEO Mark Zuckerberg's decision to not censor politicians or news (Snider, 2019). No doubt, such an article could bring into question the rest of "Big Tech" industry partners.

As such, Barrett (2020) states that aside from Section 230, a more effective means at content moderation is through transparency and improving accountability (p. 3). While Barrett (2020) proposes the advent of a regulatory agency, the concepts can be applied to how Twitter conducts its existing and future transparency practices. For example, bipartisan bill Platform Accountability and Consumer Transparency ("PACT Act") was introduced in 2020. Among other changes, the act required ISPs to "explain their content moderation policies to users and provide detailed quarterly statistics on items removed, down-ranked, or demonetized" (p. 9).

Perhaps, as a result of this, some of the top social media companies have already resorted to public announcements regarding their combative actions. By 2019, Allcott et al. (2019) noted that since the 2016 U.S. Presidential election, Facebook made up to 12 public announcements where it acknowledged misinformation on its platform and that it took steps to resolve these issues and, in contrast, Twitter had released five public announcements doing the same (p. 1). As of January 11, 2021, Twitter released their 17th Transparency report, which is aimed at "building and increasing public trust" (Twitter Inc., 2021).

However, if liability were increased towards social media companies, and there was stronger enforcement of content, a likely outcome would be that there would be less transparency in the processes (Information Society Project, Yale Law School, 2017, p. 9). Yale's Information

Society Project (2017) discussion noted that a viable solution towards combatting misinformation is through educating content consumers (p. 11). Of note, Twitter's transparency public announcements provide insight into their practices in concert with educating users regarding moderation and misinformation (Twitter Inc., 2021).

LEGAL PERSPECTIVE

Aside from misinformation, Section 230 is coming under scrutiny for a variety of reasons that are tied closely to misinformation. For example, in the case of Fields v. Twitter (2018), the Plaintiff claimed injuries as a result of Twitter Inc. providing material support to Foreign Terrorist Organization (FTO) ISIS in the form of a service. The Plaintiff supported this claimed by asserting that within one year of its [initial] proceedings in 2016, "Twitter [knowingly] allowed ISIS to attract 'more than 30,000 foreign recruits'" through propaganda an other means, which contributed to the death of two Americans abroad (Fields v. Twitter, 2018). While the plaintiff was unsuccessful in its appeal, it nonetheless demonstrated the power of Section 230 to have stopped Twitter's liability (Fields v. Twitter, 2018).

Similarly in Pennie v. Twitter (2018), the Plaintiff argued that companies Twitter Inc., Facebook, and Google provided material support (in the form of services) to FTO Hamas. The Plaintiff asserted that the platforms enabled the FTO to "radicalize and influence individuals to conduct terrorist operations outside the Middle East" through their platforms, ultimately contributing to the killing of five police officers. This case resulted in the favor of the Defendant, Twitter Inc., in part due to the immunization provided by the Communications Decency Act (Pennie v. Twitter, 2018).

While not explicit, misinformation through propaganda and its harmful effects provides the context to cases such as these. That these two cases are dismissed in part due to liability

protection afforded by Section 230 no doubt raises question regarding the efficacy of Section 230 in protecting the public good.

CONCLUSION

Based on the aforementioned discussions, it would be unlikely that given a removal of civil liability protection from Section 230, Twitter Inc. would increase its practices without creating several problems including over moderation. Rather, improvements to the user verification process and transparency practices may be more promising.

CHAPTER 3: METHODOLOGY

INTRODUCTION

This study gathered data through a mixed methods approach, integrating quantitative and qualitative data and facilitated through the use of an online survey tool. Quantitative data was gathered through a variation of a Likert scale choice responses to questions which relate to one of the three assumptions. Data processing was applied through data analytics tool, PowerBi, and responses were analyzed as an aggregate in identifying trends and popular perspectives regarding the efficacy of each assumption in the Theory of Change. Resulting from this, internal and external validity was identified. Qualitative data was obtained through the same survey in questions that allowed for freeform responses to explain one's reasoning for answering the questions. Both the quantitative and qualitative responses were mapped back to the assumptions listed in the Theory of Change to address the perceived efficacy of the assumption.

RESEARCH QUESTION AND SUBQUESTIONS

Therefore, the main research question is this: If civil liability protection is removed from Section 230, what is the perceived causal relationship between the removal of civil liability protection defined in 47 Section 230(c)(2) and its impacts on the volume of proliferation of misinformation through Twitter Inc platform. Operationally, the study attempted to measure the independent variable—the removal of civil liability protection in 47 Section 230(c)(2), if implemented—to the dependent variable—the perceived reduction in volume (or otherwise) of misinformation from the Twitter Inc. platform. This research question also attempted to identify causal and correlating outcomes that Twitter may employ as a result of the independent variable. Those sub questions include: determination of the impacts of an increase in content moderation;

determination of the impacts of an increase in user verification and authentication; and determination of the impacts of an intensification of public transparency practices.

THEORY OF CHANGE AND ASSUMPTIONS

The theory of change and assumptions for this study are as follows: If Internet Service Providers (ISPs) civil liability protection is removed (as defined by 47 USC §230(c)(2)) from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation; then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation; and then an intensification of public transparency by Twitter will significantly reduce misinformation. The assumptions behind this theory are as follows:

- Assumption 1 (A1): If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation.
- Assumption 2 (A2): If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation.
- Assumption 3 (A3): If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an intensification of public transparency by Twitter will significantly reduce misinformation.

OPERATIONAL DEFINITIONS:

For context, 47 Section 230(c)(2) states:

> "No provider or user of an interactive computer service shall be held liable on account of (A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or (B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1)."

Other operational definitions in this study include the following.

- Removal of civil liability in 47 Section 230 is defined as amendment to 47 Section 230 in which the above 47 Section 230(c)(2), 47 Section 230(c)(2)(A), and 47 Section 230(c)(2)(B) are removed from the rest of 47 Section 230.
- For the sake of this study, misinformation is defined as false and misleading content, purported in the form of authentic information and/or factual assertion(s).
- Content moderation is defined as measures by which social media companies limit and reduce user-generated content due to material defined in 47 Section 230 (c)(2)(A). Applicable content moderation for this study applies to the misinformation definition above.
- For the sake of this study, increased content moderation is defined as the adoption of moderation practices in addition to existing moderation practices through the use of algorithms, employee engagement, etc. in efforts to remove, label, warn, or limit the distribution of misinformation on the Twitter platform. An increase may also be defined as a positive trend in frequency, volume, and number of intervention of those content moderation practices.
- This study recognizes that users of ISP platforms often utilize more than one social media platform, and sometimes in concert with one another. Therefore, the term "Twitter user" refers to any individual who uses the Twitter platform, despite use of other social media platforms.
- For the purpose of this study, major Internet Service Providers are defined as top few social media companies, Twitter, Facebook, Reddit, etc. However, study for the effectiveness of proposed changed will be limited to the Twitter platform.
- For the purposes of this study, proliferation of misinformation is defined as the sharing of misinformation articles among users of the interactive computer service/social media platform through publicly available shared content internal to the platform and/or publicly available shared content external to the platform
- Significant change in proliferation of misinformation will be defined as the current quantifiable amount of misinformation compared to a future quantity of misinformation proliferation. Tentatively, a reduction in misinformation volumes of at least 10% will be considered significant.
- Verification and/or authentication of user identities is defined as the methods, practices, and processes by which social media platforms ensure that the user of such

service is authenticated and verified as the user they purport (that is, to their true identity), when signing up for such services. Verification of user identities include the use of true name and contact information for registration purposes and include authentication protocols to ensure authenticity of the identity. Further verification of user identities remove user anonymity.

- Increase in verification of user identities will be defined comparing current methods and practices of verification of user against future practices which reduce anonymity of user identities. This includes requiring quantitatively more and/or all users to be "verified" and/or "authenticated."
- Interactive computer service is defined as internet-based services where users may register, communicate with other users utilizing the service, upload digital content, and view digital content. Further, the term "interactive computer service" means 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, according to 47 Section 230. For the sake of this study, "interactive computer service" may be synonymous as social media company and services the social media company provides.
- For the sake of this study, public transparency announcements refers to the practice by which social media companies identify actions taken, business practices enforced, and results of those actions regarding efforts to combat misinformation.
- Intensifying public announcements is defined as at least one additional public announcement per year in which social media companies release public announcements regarding combative actions against misinformation. Intensifying public announcements also includes an increase of quantifiably distinct topics discussed in misinformation transparency announcements.

POPULATION SAMPLING STRATEGY AND PROCEDURE:

Population Sampling. Population sampling was conducted through an online survey distributed through a variety of social media platforms, email communications, and personal communications and attempted to reach at least 100 recipients to gain both quantitative and qualitative data. Public surveys will were made available to Twitter users, irrespective to any demographic qualifications or limitations. A Twitter profile advertising the intent of the research was created and utilized to share the online survey, however, this provided negligible results. The Twitter profile used for polling is listed in the Appendix.

Subject Matter Experts (SMEs). The following individuals were selected and interviewed for their subject matter expertise and provided additional qualitative responses:

- **Niam Yaraghi (2021).** He is an assistant professor of Business Technology at Miami Herbert Business School at the University of Miami and fellow at the Brookings Institution's Center for Technology Innovation, provided additional context to Twitter's moderation. N. Yaraghi has been featured on U.S.news.com and wrote several professional articles including an article entitled, "How should social media platforms combat misinformation and hate speech?" (N. Yaraghi, 2019).
- **Matthew Benassi (2021).** He and his wife, Maatje Benassi, were victims of a widespread conspiracy theory surrounding the origin of the COVID-19 pandemic as Maatje was falsely labeled as "coronavirus patient zero" by a prominent conspiracy theories and were featured in a CNN Business exclusive article, titled "Exclusive: She's been falsely accused of starting the pandemic. Her life has been turned upside down" (O'Sullivan, 2020). Further, Benassi (2020) provided perspectives on Section 230 and its impacts on social media companies and victims of similar issues.

A list of questions to the following SMEs is listed in the Appendix.

DATA PROCESSING AND ANALYSIS

Responses from surveys quantitatively measured the responded levels of agreement to potential outcomes as a result of the independent variable. Analysis was performed on the conglomerate of all responses in order to determine common popular perceptions. Analysis included the following: the most common responses to specific questions; identification of the most frequently agreed-upon result; relationships between answers among users.

Freeform responses to the survey questions provided qualitative perspective to the study and were compared to the survey-based analysis. The comparison of freeform responses from the surveyed combined with radio-button responses provided a multi-dimensional perspective towards the efficacy of implementing each assumption.

This information subsequently mapped back to one of the three assumptions identified in the Theory of Change.

INTERNAL AND EXTERNAL VALIDITY AND LIMITATIONS

A potential benefit of utilizing the Twitter poll tools to conduct survey is that users would be able to conduct this survey anonymously, while in any setting they so choose and already be a Twitter user. Because this method yielded negligible results, an online survey tool was utilized across multiple forms of communication to engage respondents. Therefore, an additional qualifying question was question asked of respondents: "Have you ever used and/or referenced information that came from Twitter in any capacity?" This resulted in 64% of respondents answering "Yes," therefore potentially limiting the internal validity of the responses provided by the remaining 36%. While it could be argued there is reduced credibility in respondents who answered "No" for this question, it also could be argued that there are applicable reasons as to why those respondents do not reference information from Twitter. For example, these respondents may or may not favor Twitter's public perception.

Without the presence of face-to-face interaction, the surveyed are subject to their own interpretation of the question, no matter the degree of specificity of how the questions are presented. These interpretations may be influenced by the surveyed users' own personal bias, worldview, language barrier, skepticism, additional input of perceived factors, etc. There will likely be no opportunity for one to be present to help clarify intent of questions or otherwise. For

example, 36% of respondents answered “No” when asked the qualifying question, “Have you ever used and/or referenced information that came from Twitter in any capacity?” However, some respondents who answered “No” could have actually answered “Yes” after feedback discussion regarding the survey.

Further, online survey was conducted for a limited time frame of no more than three weeks, thereby significantly limiting the population of subjects to be surveyed.

Regarding A1, this study recognizes that the potential change of removing civil liability protection in 47 USC §230 applies civil liability to ISPs, but not the users themselves. While it can be argued that the legal liability may arguably fall upon the creator of the content (e.g. the user) and not the distributor, it can also be argued that social media companies like Twitter will continue to moderate content on its platform.

Regarding A2, this study recognizes that it does not speak on behalf of Twitter Inc., nor can it provide insight to internal business decisions. However, it is worth studying the perceived efficacy of improvements to user identity verification processes and potential impacts to user behavior as a result of the independent variable. It is worth studying the potential impact of the reduction in anonymity upon user experience and retention.

Regarding A3 and as stated regarding A2, this study does not speak on behalf of Twitter Inc., nor can it provide insight to internal business decisions. However, it is worth studying the perceived efficacy of potential intensification of Twitter’s public announcements as a result of the independent variable.

The information gained from this study may be applicable to other social media companies, particularly those concerned about moderating misinformation. This study will also solicit ideas from users about how to tackle the misinformation volume problem.

SUMMARY

In summary, while efforts are made to present objective and clear wording and presentation of surveys, there are limitations to the data collected in addition to the validity of the data. Therefore, the combination of surveys and potential interviews of SMEs provide a multi-dimensional understanding of the effects of the independent variable upon each assumption.

CHAPTER 4: RESULTS AND FINDINGS

Introduction

This chapter presents the results, findings and analyses of the qualitative and quantitative data collected using an online survey with 97 respondents and completion of subject matter expert interviews. Each of the questions in the survey were multiple choice; a few of the questions offered an additional opportunity for the respondents to provide qualitative responses to the questions. With an initial goal of 100, 97 respondents participated in the survey. Respondents provided a total of 150 qualitative responses in addition to over 660 multiple choice answers responses.

ASSUMPTION 1 FINDINGS

Assumption 1 (A1) is restated as follows: If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation. Survey Question 3 asked, "Currently, Twitter moderates misinformation by removing tweets/account, flagging, limiting exposure, etc. If Twitter increased misinformation measures, then misinformation on Twitter will…"

A1 Quantitative Results Validation and Findings: Regarding Survey Question 3, overwhelmingly, 59% of respondents reported that an increase in moderation for misinformation content by Twitter would result in a decrease of misinformation on the Twitter platform by 10% or more. However, a substantial 31% of respondents believed this change would result in negligible impacts upon misinformation volumes. Few remaining respondents believed an increase in moderation would cause more 10% or more misinformation on Twitter. Reference Figure 1.

A1 Quantitative Alignment to Literature Review: The literature review by scholars and subject matter experts surrounding an increase moderation generally suggested that an increase in moderation would produce negligible impacts to reducing volumes of misinformation in addition to potentially negative tertiary consequences such as "over moderation." However, respondents largely contrast this viewpoint, which suggests that respondents believe Twitter is able to reduce misinformation by at least 10% through increasing content moderation measures.

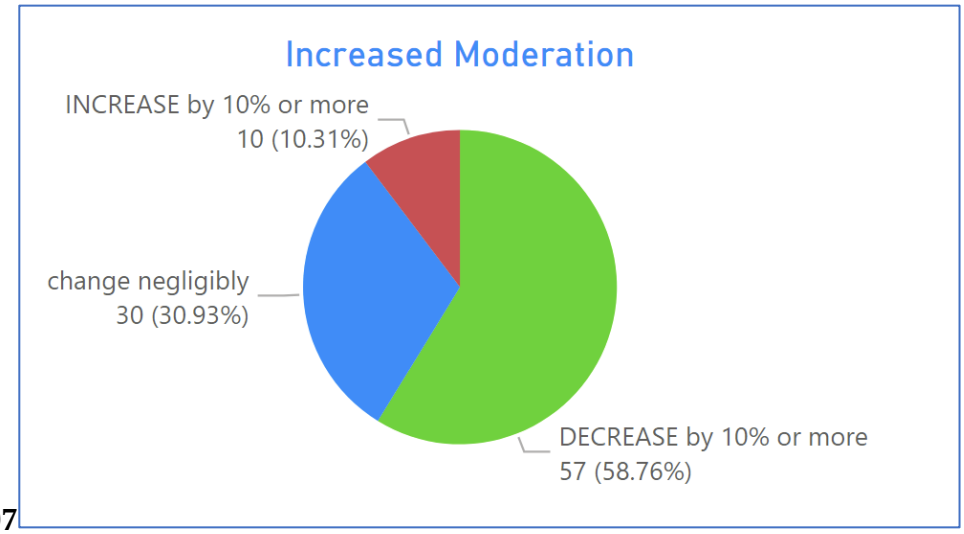

Figure 1: N=97

A1 Qualitative Survey Results Validation and Challenges: Of 21 respondents that provided open ended responses, at least five respondents shared concern surrounding Twitter's existing moderation practices, therefore demonstrating further concern regarding an increase in moderation by the company.

Of respondents who responded that an increase in moderation would result in decreasing misinformation by 10% or more, respondent provided a wide variety of amplifying detail. One respondent stated that "Twitter seems to moderate in one direction only. Cracking down may decrease misinformation, but it may also silence true information…"

Of respondents who believed an increase in moderation would result in negligible change to misinformation on Twitter, few respondents suggested that Twitter may be challenged in retaining its userbase. For example, respondents stated, "censorship would certainly lead to Twitter user churn, and even negatively impact the platform" and that users will "therefore go to other sites." One responded stated "If social media is held legally liable for the mis-information [sic] shared by individuals they will drastically limit free speech In [sic] an attempt to remain legally bullet proof [sic]."

Of respondents who believed such an action would cause an increase in misinformation by 10% or more, most respondents shared concern regarding the integrity of moderation practices by Twitter Inc., alluding to a perception of the company's political bias.

A1 Interview Results Validation and Challenges: N. Yaraghi states that social media companies are placed in a challenging situation as they were gifted a "white elephant" of content moderation, although this was not their initial ambition. Social media companies started out by trying to provide a medium of social connections among users, however they have now been reluctantly placed in a position in which they are forced to moderate content because it has been recently recognized that their platforms can influence important issues like the results of an election (N. Yaraghi, personal communication, May 18, 2021).

Yaraghi references the "What's Happening" page on a Twitter login screen, which highlights various tweets. As a result of such a module, Yaraghi states that Twitter distances itself from being a pure platform and closer to that of a publisher. It further demonstrates Twitter is quite capable of moderating content on its platform (N. Yaraghi, personal communication, May 18, 2021).

Yaraghi identifies an issue in a hypothetical situation in which all content was moderated on Twitter. If this were the case, the situation would change Twitter into a platform like Forbes in which all information is verified, thereby changing its business model. However, the problem that would remain is that it would still not be unbiased. As it is today, content on Twitter has a lot of opinions and is different than an organizations like the New York times, which provides content that can be verified (N. Yaraghi, personal communication, May 18, 2021).

When asked "If Social Media companies like Twitter undertook one of the following, which would be of the most benefit to users like you to mitigate or address spread of misinformation. a. Increase moderation on user-generated content against misinformation b. Verify/Authenticate user profiles to identities of it users c. Intensify company transparency announcements on moderation of misinformation issues, including educating users on misinformation topics, issues, and events," M. Benassi responded:

> Because our main harasser hasn't tried to hide I would say that a. increase moderation on user-generated content against misinformation would've helped us the most. But honestly they will not do this because of the CDA Section 230. Moderating anything more than the bare minimum would hurt their revenue stream and the law provides them nearly absolute [sic] no recourse from most victims. For this reason, the only real solution to this is to modify Section 230 to give protection to victims such as our selves. (M. Banassi, personal communication, May 21, 2021).

Additionally, M. Benassi's response to other questions can help provide further color to A1 in its relation to user-generated content moderation. He states that currently, Twitter does not "see any issue" with the harasser's account as "he hasn't violated their terms of service" (Benassi, 2021). Benassi further acknowledges that a censoring or removal of the harasser's

account would make it challenging to "reconstitute his network very easily—he would lose his funding streams and therefore his ability to connect with his followers. Twitter needs to look at the bigger picture and not just what is happening on its platform" (Benassi 2020). He further stated that "if Section 230 were modified to force social media companies to limit harassment and defamation of innocent victims, that would be a good thing…" (M. Banassi, personal communication, May 21, 2021).

ASSUMPTION 2 FINDINGS

Assumption 2 (A2) is restated as follows: If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation.

A2 Quantitative Survey Results Validation and Challenges: Survey Question 4 was asked, "Currently, Twitter verifies/authenticates users, but many still use anonymous/fictitiously-named profiles. If Twitter required verified/authentic profiles, then MISINFORMATION on Twitter will...". As referenced in Figure 2a, overwhelmingly 63% of respondents believed misinformation would decrease by 10% or more on the Twitter platform. A substantial amount of respondents, 31%, believed that it misinformation volumes would change negligibly, and remaining few believed misinformation would increase by at least 10% or more on Twitter.

Also, results from an additional survey question provide additional dimension to A2. Survey Question 6 asked, "If Twitter undertook one of the following to fight misinformation (all in consideration of Freedom of Speech, Privacy, User Experience, etc.), which would be of MOST BENEFIT to users?." Of the responses, a substantial 52% of respondents indicated that having verified and authenticated users on Twitter would be of most benefit to users, as

compared to Twitter implementing an increased moderation or intensifying transparency, as referenced in Figure 2b.

A2 Quantitative Alignment to Literature Review: The majority of responses indicate alignment with the literature review concerning the application of verified and authenticated users. The literature review generally suggested that combinations of having verified and authentic users—such as Twitter's Blue Verification badge—was hopeful endeavor in having a user base that engages in information verification practices and behaviors.

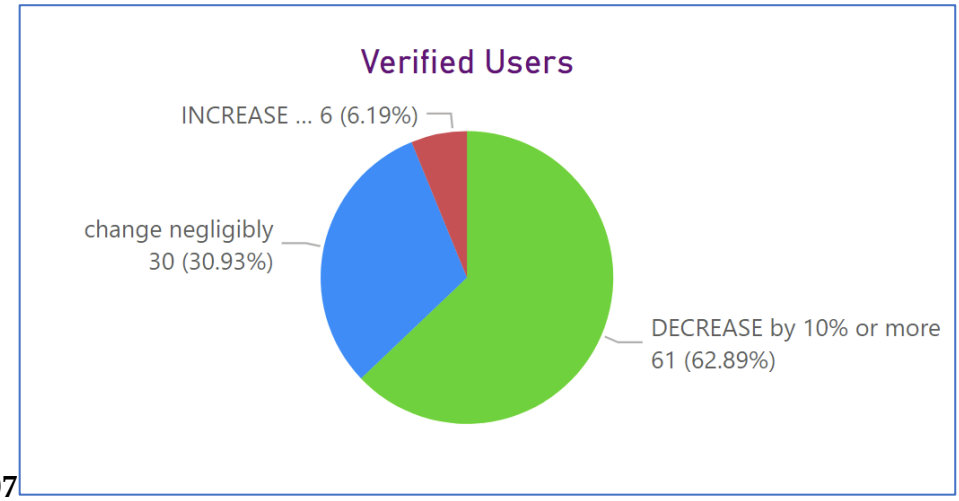

Figure 2a: N=97

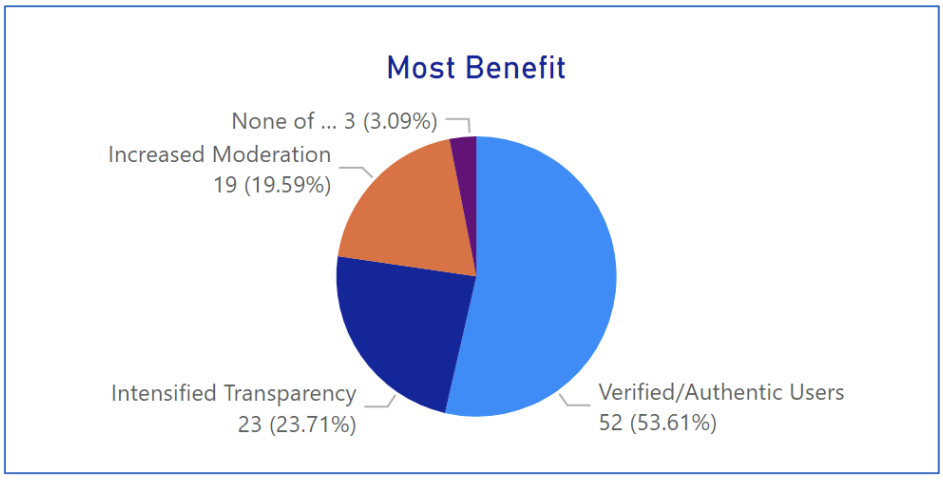

Figure 2b: N=97

A2 Qualitative Survey Results Validation and Challenges: From the responses of Survey Question 4, there were 17 qualitative responses provided.

Of users who believed that if Twitter required users to have verified and authentic profiles, users provided variety of responses, but many discussed issues surrounding user anonymity. Respondents conveyed levels of anonymity would be reduced, thereby potentially affecting user privacy. For example, having a verified account would allow a user to have "credibility when providing information" or would limit "trolls / fake accounts" which are often "used for propaganda or fake news or misinformation." In spite of their response, a respondent acknowledged that "creating a dummy account is a freedom that everyone is entitled" and that they may not be willing to divulge their personal information.

Of respondents who believed that there would be negligible change, one respondent acknowledged the challenge to both users and Twitter, stating "the user experience would shift dramatically," and that the cost to Twitter would be extensive, while this would still not eliminate the problem of "bots and fakes." Another respondent acknowledged the global issue where verification methods may differ from country to country.

A2 Interview Results Validation and Challenges: N. Yaraghi stated that there could be substantial negative consequences to the Twitter user experience if it required verification of its users. This would increase pressure on users, which would cause users—that is, those who prefer communication through "pseudonames" and express themselves without cost—to pursue or migrate to another platform. Consequently, rules and regulations only work if users can be held accountable, therefore users may be incentivized to move to another platform if Twitter required users to be verified and/or authentic (N. Yaraghi, personal communication, May 18, 2021).

As stated previously, M. Benassi was harassed by a user who "hasn't tried to hide" and uses his real name. However, other accounts associated with proliferating the conspiracy or

harmfully reacting to the conspiracy were utilizing “fictitious names” on various social media platforms. M. Benassi did not make an assessment on whether users should be authenticated and verified, however he detailed how fictitiously-named accounts created challenges for him:

> On Twitter the individual who doxxed [sic] us and identified my parents was using a fictitious name and also on YouTube it appears the accounts that were making death threats were using fictitious names as well. If we considered bringing civil litigation against any of those parties, having their real names would make that a much easier process. M. Benassi (personal communication, May 21, 2021)

A2 Qualitative Alignment to Literature Review: In general, qualitative insights provided perspectives not extensively noted in the literature review. The benefit provided by the qualitative responses is a unique perspective from the user and impacts to user experience. Respondents provided insight into user experience which may or may not desire anonymity. Both survey and interviews provide company perspectives which indicate high costs for the company.

ASSUMPTION 3 (A3) FINDINGS & CONCLUSION:

Assumption 3 (A3) is restated as follows: If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an intensification of public transparency by Twitter will significantly reduce misinformation.

A3 Quantitative Results Validation and Challenges: Survey Question 5 asked, “Currently, Twitter shares transparency reports on moderation. If Twitter INTENSIFIED transparency reporting (e.g., publicize algorithms; educate users on more misinformation topics; etc.), then MISINFORMATION on Twitter will…”. As a result of intensified transparency on

moderation, Figure 3a shows that overwhelmingly, 66% of respondents believed that this would decrease misinformation on Twitter by 10% or more. This is followed by 27% of respondents who believed it would cause negligible change in volumes of misinformation, and a remaining 7% believed it to cause an increase of 10% or more of misinformation on Twitter.

Also, results from an additional survey question provide another dimension to A3. Survey Question 7 asked, "If Twitter undertook one of the following to fight misinformation (all in consideration of Freedom of Speech, Privacy, User Experience, etc.), which would be of LEAST BENEFIT to users?" The results of this question indicated that an Intensified Transparency by Twitter would be least beneficial to users, as compared to Twitter implementing an increased moderation or verifying and/or authenticating users, as referenced in Figure 3b. However, the margins differed by only a few votes.

A3 Quantitative Results Alignment to Literature Review: The literature review gives a sense of positive benefit to the application of improved transparency practices. While transparency does not directly apply actions against content or content creators, it provides benefits of education, topical discussion, and boosts company perception. In general, the quantitative results associated with A3 support the literature review.

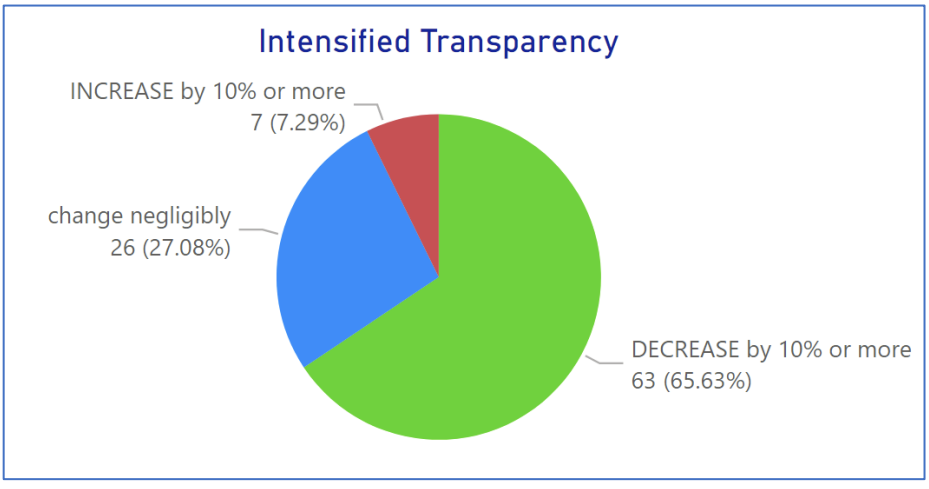

Figure 3a: n=96

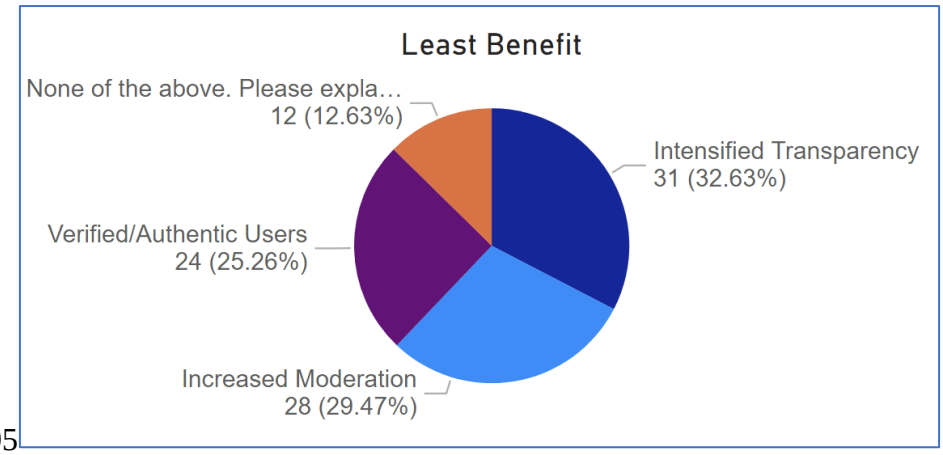

Figure 3b: n=95

A3 Qualitative Survey Results Validation and Challenges: Regarding Survey Question 5, there were 13 qualitative responses provided in addition to the multiple choice selection.

Of respondents who believed that intensifying transparency would result in a decrease of misinformation by 10% or more, respondents commented that it would "create awareness to everyone on the legitimacy of the information," and "transparency forces everybody to be more responsible…imagine your true identity will be out because of spreading false information."

Respondents that believed that intensifying transparency would result in negligible results and those who believed it could cause a 10% or more increase in misinformation stated that "publicizing transparency moderation process might provide loopholes to users." Others casted doubt the political bias and integrity of Twitter, associating the platform with an ideological agenda.

A3 Interview Results Validation and Challenges: While N. Yaraghi did not have comments regarding Twitter's transparency practices, his reference to Twitter's relationship with other social media companies may provide applicable insights. Yaraghi mentioned that an increase in moderation by Twitter would not be an issue, for example deactivating a perpetrator's account. However, social media companies have demonstrated that they may often work

together to deactivate or disrupt services for an individual, which causes significant hardship for that person to get services elsewhere. Analogously, if an airline were to decline a customer transportation services, that customer may be able to get service at another airline. However, if all the airlines colluded with one another and all declined service to that individual, then this would create significant hardship for an individual to travel long distances at all. Something similar happened to the Parler application. In this day and age—especially in light of COVID-19—a person's online presence may be more important than their physical presence (N. Yaraghi, personal communication, May 18, 2021).

As Twitter is often named among "Big Tech" companies, this type of collusion may no doubt cause a negative perception upon Twitter's perceived political bias.

M. Benassi provided insight into the importance of transparency as it pertains to potential reform:

> Clearly, harassment/misinformation/defamation runs rampant on all social media platforms and if Twitter actually came clean about the volume of that on their platform, then we would probably get some real movement on Section 230 reform. It would be hard to ignore. M. Benassi (personal communication, May 21, 2021)

A3 Qualitative Results Alignment to Literature Review: Qualitative insights vaguely support the viewpoints of the literature review. For example, respondents generally acknowledged that there is not direct effort against combatting misinformation itself. However, an increase in transparency and metrics can only help the company, but Twitter also has an uphill battle regarding perception of its political leanings.

CHAPTER 5: CONCLUSIONS, RECOMMENDATIONS, & AREAS FOR FURTHER RESEARCH

Overall, the proposed Theory of Change (TOC)—that is, if Internet Service Providers' (ISPs) civil liability protection is removed (as defined by 47 USC §230(c)(2)) from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation; then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation; and then an intensification of public transparency by Twitter will significantly reduce misinformation—is generally validated. Validation is limited by the perceived efficacy of each assumption provided by respondents and interviewees but does not reflect actual outcomes if such actions are implemented. At a high level, results garnered from this research identify user perspectives and should inform Twitter Inc. of potential challenges and benefits as a result of implementing any of the aforementioned actions in the TOC. While each assumption was generally supported by quantitative data, qualitative data often challenged the efficacy of each assumption. For example, Assumption 2 (A2), which was perceived as most beneficial to users as compared to other assumptions, may come at significant cost to both users and the company. Additionally, qualitative data suggests that Twitter Inc. suffers perception of being politically biased, and therefore any actions by Twitter Inc. to correct misinformation issues will likely be accompanied by a substantial public suspicion.

Assumption 1: If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation.

Assumption 1 Conclusion: Quantitative responses suggest that increasing moderation will be mostly an effective measure to significantly reduce misinformation (that is, to reduce misinformation on Twitter by 10% or more). However, qualitative insights indicate concern over Twitter's existing moderation practices and challenges to user experience if such a measure is implemented. An increase in moderation would potentially cause a significant change in its business model, and therefore changing user experience.

Assumption 2: If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation.

Assumption 2 Conclusion: Quantitative results generally suggest that requiring Twitter users to be verified and authenticated would mostly be effective at significantly reducing misinformation (that is, to decrease misinformation on Twitter by 10% or more). Respondents perceived A2 to be significantly more effective against misinformation than other assumptions. However, qualitative results suggest that this would come at great cost to the company and user experience. The market for user anonymity cannot be ignored, and the cost for Twitter Inc. may be significant. However, user anonymity may not be necessarily correlating to spreading misinformation because verified, non-anonymous accounts may be perpetrators of widespread misinformation.

Assumption 3: If Internet Service Providers (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an intensification of public transparency by Twitter will significantly reduce misinformation.

Assumption 3 Conclusions: Quantitative results suggest that intensifying transparency is mostly beneficial against significantly reducing misinformation (that is, to decrease misinformation on Twitter by 10% or more). Quantitative results suggest that intensifying transparency is marginally the least benefit to users in its attempt to significantly reducing misinformation, as compared to other assumptions. Qualitative data suggests that user awareness and public accountability for both Twitter Inc. and perpetrating users is beneficial. Qualitative insights indicate that Twitter Inc. suffers a public perception of harboring political bias, despite its existing transparency practices.

RECOMMENDATIONS:

Regarding A1—that is, if Internet Service Providers' (ISPs) civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an increase of content moderation practices by Twitter will significantly reduce misinformation—and A3—that is, if ISPs' civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then an intensification of public transparency by Twitter will significantly reduce misinformation—Twitter Inc. should…

- …update its tweet process allowing users to include additional category selection for their tweet, for example "fact," "opinion," "speculation" or something similar. If the user chooses to categorize their tweet as a "fact," then require the user link a URL to the source of information, or else allow the user to default it to an "opinion" or similarly-labeled category. Twitter may generate a Database of Reputable and Reliable Sources of Information, for example mainstream news outlets, peer reviewed journals, etc., to which the URL may link. The "fact" category must map to the domain of one of those organization listed in the

Database of Reputable and Reliable Sources of Information. Organizations wishing to be part of this database may undergo a certification, screening, verification, fact check process, and may regularly be recertified to be included on that database. This may assist in Twitter's algorithm for sharing and moderation, and the database may be included in the Twitter's public transparency report. Implementation of this effort may be somewhat long-term and should be projected to be implemented with the appropriate software updates and advertisement of its certification process by at earliest end-of-calendar year 2022.

- …publicly acknowledge concerns surrounding its perceived political bias, regardless of the legitimacy accusations against the company. Improvements to its moderation will be met with challenges if perceptions against Twitter Inc. go unaddressed. Twitter may provide this acknowledgement through advertisement of its policies and public announcement through highly visible feeds, such as the "What's Happening" page, which may provide an additional link to its public transparency announcements. This should be accomplished as soon as possible as to mitigate further damage to Twitter's reputation.

Regarding A2—that is, if ISPs' civil liability protection is removed from Section 230 to incentivize reduction of misinformation on social media platforms, then increase in user identity verification and authentication policies by Twitter will significantly reduce misinformation—Twitter Inc. should increase its user verification and authentication practices with the following caveats. Twitter should utilize identity applications such as ID.me to allow users to sign into Twitter application with an authentic/verified identity but be allowed to communicate with pseudonym if they choose. This would provide privacy to communicate widespread, while

accountability to Twitter policies and law enforcement. This would require development of a relationship with an existing application that provides such services, such as ID.me. This will be achievable as it will lessen the administrative burden of PII from being contained on the Twitter platform. However, software update and integration with a third party will be required, therefore making this a long-term project. Twitter may beta test this process with the roll out by end of the next calendar year, 2022.

AREAS FOR FURTHER RESEARCH:

There are several areas for further research due to time and resource limitations but are worth entertaining for further research:

- Low-cost solutions for user identify verification and authentication for social media. Various open-source articles allude to blockchain technologies.
- From a legal perspective, the feasibility of civil ability if civil liability protection is removed from Section 230.
- Other potential changes to Section 230, such as redefining various terminology.
- User's current understanding and familiarity of Twitter's existing transparency reports.
- Impacts of propaganda, disinformation, and state-sponsored foreign influence as it pertains to targeting and exploitation on social media platforms.
- Respondents' general knowledge of volumes of information and misinformation, computer and data sciences, and legal implications, which may contribute to a contrast with the literature review.

REFERENCES

2017 Practicum Research Team (2017). Fake News & Misinformation Policy Practicum. Stanford Law School, Law and Policy Lab. Retrieved from http://law.stanford.edu/policy-lab.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. Research & Politics, *6(2)*, 1-8. https://doi.org/10.1177/2053168019848554

Armijo, Enrique (2021). Reasonableness as Censorship: Section 230 Reform, Content Moderation, and The First Amendment. Florida Law Review, Forthcoming 73, 1-37. https://ssrn.com/abstract=3764061.

Barrett, Paul (2020). Regulating Social Media: The Fight Over Section 230—and Beyond. NYU STERN Center for Business and Human Rights, 1-18. https://blog.sodipress.com/wp-content/uploads/2020/09/regulating-social-media.pdf

Butler, A. (2018). Protecting the Democratic Role of the Press: A Legal Solution to Fake News. Washington University Law Review, 96(2), 419-440. https://heinonline.org/HOL/LandingPage?handle=hein.journals/walq96&div=15&id=&page=

Benassi, Matthew (2020). A conspiracy theory almost ruined my family's life. This could prevent it from happening to you. CNN Business Perspectives. https://www.cnn.com/2020/10/23/perspectives/section-230-disinformation-hate-speech-social-media/index.html

Cheah, M. A. (2020). Section 230 and the Twitter Presidency. Northwestern University Law Review Online, 115, 192-222. https://heinonline.org/HOL/LandingPage?handle=hein.journals/nulro115&div=9&id=&page=

Cramer, B. (2020). From Liability to Accountability: The Ethics of Citing Section 230 to Avoid the Obligations of Running a Social Media Platform. Journal of Information Policy, 10, 123-150. doi:10.5325/jinfopoli.10.2020.0123

Fields v. Twitter, Inc., 881 F.3d 739 (9th Cir. 2018). Retrieved from https://casetext.com/case/fields-v-twitter-inc-3?PHONE_NUMBER_GROUP=P&sort=relevance&resultsNav=false&tab=keyword

Goodman, E.P., & Wittington, R. (2019). Section 230 of the Communications Decency Act and the Future of Online Speech. The German Marshall Fund of the United States, 20, 1-12. https://www.jstor.org/stable/pdf/resrep21228.pdf

Goodyear, Michael (2020). Is There No Way to the Truth? Copyright Liability as a Model for Restricting Fake News. Harvard Journal of Law & Technology, 34, 279-306. http://dx.doi.org/10.2139/ssrn.3647504

Gupta, L., Gasparyan, A. Y., Misra, D. P., Agarwal, V., Zimba, O., & Yessirkepov, M. (2020). Information and Misinformation on COVID-19: a Cross-Sectional Survey Study. Journal of Korean medical science, 35(27), e256. https://doi.org/10.3346/jkms.2020.35.e256

Klein, I. (2020). Enemy of the People: The Ghost of the F.C.C. Fairness Doctrine in the Age of Alternative Facts. Hastings Communications and Entertainment Law Journal, 42(4), 43-76. https://repository.uchastings.edu.

Lee, T. D. (2021). Combating Fake News with 'Reasonable Standards.' Hastings Communications and Entertainment Law Journal, 43(4), 81-108. https://repository.uchastings.edu.

Levush, R. (2020). Government Responses to Disinformation on Social Media Platforms: Comparative Summary. Library of Congress. https://www.loc.gov/law/help/social-media-disinformation/compsum.php

Office of the Attorney General, Department of Justice (2020). Department of Justice's Review of Section 230 of The Communications Decency Act of 1996. Department of Justice Archives. https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996.

O'Sullivan, D. (2020). Exclusive: She's been falsely accused of starting the pandemic. Her life has been turned upside down. CNN. https://www.cnn.com/2020/04/27/tech/coronavirus-conspiracy-theory/index.html

Ognyonova, K., Laser, D., Robertson, R., Wilson, & Christo, W. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School, 1(4), 1-19. https://doi.org/10.37016/mr-2020-024

Pennie v. Twitter, Inc., 281 F. Supp. 3d 874 (N.D. Cal. 2017). https://casetext.com/case/pennie-v-twitter-inc-1?PHONE_NUMBER_GROUP=P&sort=relevance&q=Pennie%20v.%20Twitter&p=1&type=case

Rojo, Camino, Twitter Inc. (2021). #SafterInternetDay 2021: Together for a better Internet. Twitter Inc. https://blog.twitter.com/en_us/topics/company/2021/saferinternetday2021togetherforabetterinternet.html

Snider, Mike (2019). Facebook's Mark Zuckerberg says the social network should not be 'censoring politicians.' USA Today. https://www.usatoday.com/story/tech/talkingtech/2019/12/02/mark-zuckerberg-facebook-should-not-censor-politicians-ads/4350547002/

Torres, R. R., Gerhart, N., & Negahban, A. (2018). Combatting Fake News: An Investigation of Information Verification Behaviors on Social Networking Sites. Hawaii International Conference on System Sciences 2018: Truth and Lies: Deception and Cognition on the Internet, 3976-3985. Retrieved from http://hdl.handle.net/10125/50387

Twitter Inc. (2021). Insights from the 17th Twitter Transparency Report. https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html.

Twitter Inc. (2020). Our plans to relaunch verification and what's next. https://blog.twitter.com/en_us/topics/company/2020/our-plans-to-relaunch-verification-and-whats-next.html

Twitter Inc. (2020). About verified accounts. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Twitter Help Center, Twitter Inc. (2021). The Twitter Rules. https://help.twitter.com/en/rules-and-policies/twitter-rules.

Twitter Safety, Twitter Inc. (2019). Disclosing new data to our archive of information operations. https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.html.

Verstraete, M., Bambauer, D. E., Yakowitz B., Jane R. (2021) Identifying and Countering Fake News. Hastings Law Journal, 73, 1-39. http://dx.doi.org/10.2139/ssrn.3007971

The Information Society Project, Yale Law School (2017). Fighting Fake News Workshop Report. Yale Law School and the Floyd Abrams Institute for Freedom of Expression. Retrieved from

https://law.yale.edu/sites/default/files/area/center/isp/documents/fighting_fake_news_-_workshop_report.pdf.

Yaraghi, N. (2019). How should social media platforms combat misinformation and hate speech? Brookings. https://www.brookings.edu/blog/techtank/2019/04/09/how-should-social-media-platforms-combat-misinformation-and-hate-speech/

APPENDIX A: Twitter Poll Account and Survey Questions

- Name / handle: @Section230Poll
- Bio: To gain survey data regarding changes
- Non affiliation statement: This Twitter account is not affiliated with Twitter Inc and is not affiliated with any political party or opinion. Additionally, polling content and questions do not support, nor defend, the removal of civil liability protection from Section 230. This poll is used for academic purposes in order to gauge Twitter user's perspectives on the efficacy of removing civil liability protection and its potential impacts upon Twitter.
- Other profile disclaimer info

Survey Questions:

1) Have you ever used and/or referenced information that came from Twitter in any capacity?
    a. Yes
    b. No
2) In light of rampant social media misinformation, will removing liability protection from Section 230 (i.e. making social media companies open to liability for content) cause Twitter to take MORE action to reduce misinformation?
    a. Yes (or probably)
    b. Neutral
    c. No (or probably not)
3) Currently, Twitter moderates misinformation by removing tweets/accounts, flagging, limiting exposure, etc. If Twitter INCREASED misinformation moderation measures, then MISINFORMATION on Twitter will...
    a. Decrease by 10% or more
    b. Change negligibly
    c. Increase by 10% or more
4) Currently, Twitter verifies/authenticates users, but many still use anonymous/fictitiously-named profiles. If Twitter required verified/authentic profiles, then MISINFORMATION on Twitter will...
    a. Decrease by 10% or more
    b. Change negligibly
    c. Increase by 10% or more
5) Currently, Twitter shares transparency reports on moderation. If Twitter INTENSIFIED transparency reporting (e.g. publicize algorithms; educate users on more misinformation topics; etc.), then MISINFORMATION on Twitter will…
    a. Decrease by 10% or more
    b. Change negligibly
    c. Increase by 10% or more
6) If Twitter undertook one of the following to fight misinformation (all in consideration of Freedom of Speech, Privacy, User Experience, etc.), which would be of MOST BENEFIT to users? After choosing, please comment why you chose it.
    a. Increased Moderation

   b. Verified/Authentic Users
   c. Intensified Transparency
   d. None of the above. Please explain.

7) If Twitter undertook one of the following to fight misinformation (all in consideration of Freedom of Speech, Privacy, User Experience, etc.), which would be of LEAST BENEFIT to users? After choosing, please comment why you chose it.
   a. Increased Moderation
   b. Verified/Authentic Users
   c. Intensified Transparency
   d. None of the above. Please explain.

APPENDIX B: Subject Matter Expert Interview Questions

Questions for Niam Yaraghi:

What would be the consequences (costs and benefits) if Twitter increased its current moderation practices in efforts to address as much misinformation as possible?

How effective is this in moderating misinformation such as propaganda from state-sponsored groups, and/or bots?

A university study recommended that all users be allowed to gain access to the blue verification badge.

What are your thoughts on its effectiveness towards mitigating misinformation?

Twitter currently has a pages dedicated to statistics and reports on what it moderates.

Are you familiar with this? If so, how effective is public transparency (i.e. reports, education on misinformation, etc.) to mitigating and reducing misinformation?

Questions for Matt Benassi:

1)Which social media company(s) were involved in the development of misinformation spread about you and/or your family?

2)(General Qualifying questions) What type of misinformation was spread about you and/or your family member(s)?

a.___. Defamation

b.___. Propaganda

c.___. Conspiracy

d.___. Disinformation

e.___. Other(s): _____

3)What were the general impacts to you and/or your family regarding the challenges caused by misinformation? For example, social, financial, professional, etc. (Please don't feel compelled to provide details. A general list will do, only if you are comfortable).

4)If Social Media companies like Twitter undertook one of the following, which would be of the most benefit to users like you to mitigate or address spread of misinformation.

a. Increase moderation on user-generated content against misinformation

b. Verify/Authenticate user profiles to identities of it users

c. Intensify company transparency announcements on moderation of misinformation issues, including educating users on misinformation topics, issues, and events

d. Others: _____

5)**If social media companies like Twitter were to have increased moderation against misinformation on their platform (i.e. flagging of misinformation content, removing messages and/or accounts involved in proliferating misinformation, reducing exposure of information, etc.)…

a.…how would this have impacted you and/or your family in terms of mitigation and severity of issues caused by misinformation?

b.…what would be other implications to the company and users? For example, impacts to freedom of speech, defamation, etc.

6)**What would be the implications if social media companies like Twitter implemented policies and processes that required user profiles to be authentic and verified to the user's actual identities? For example, please consider freedom of speech, defamation, anonymity, privacy, user experience, and other issues.

a. If applicable and regarding the individual(s) spreading false and misleading information about you and your family, were the Individuals Spreading Misinformation using authentic or fictitious/alternate personas/profiles (i.e. under another name, alias, unverified account)?

b. If applicable and regarding the individual(s) spreading false information and misleading about you and your family, how difficult/easy was the process of identifying the Individuals Spreading Misinformation based on information contained in their account/profiles?

7)If social media companies like Twitter were to have included topics/issues/controversies such as yours in its public transparency report(s), how would that have affected the prevention or development of challenges you experienced with misinformation?

a. What do you believe was the reason your misinformation challenges were/were not covered in a social media company's transparency report regarding misinformation?

b.**What would be the impacts if social media companies like Twitter intensified its transparency reports (e.g. to include moderation misinformation issues such as misinformation spread about you and/or your family, educating users on misinformation topics on its platform, moderation statistics, etc.).

8)If civil liability protection was removed from Section 230 allowing Internet Service Providers (ISPs) (such as social media companies to become more liable for content on their platform), how would that have impacted the development or mitigation of challenged you and/or your family experienced as a result of misinformation? (For example, would it have mitigated its severity or made it easier to resolve?)