California Law Review

Volume 106
Issue 6 *12/2018*

Article 3

12-1-2018

#I?U: Considering the Context of Online Threats

Lyrissa Barnett Lidsky

Linda Riedemann Norbut

Recommended Citation

Lyrissa Barnett Lidsky and Linda Riedemann Norbut, *#I?U: Considering the Context of Online Threats*, 106 Calif. L. Rev. 1885 (2019).

Link to publisher version (DOI)
10.15779/Z38JM23G4C

#I  U: Considering the Context of Online Threats

Lyrissa Barnett Lidsky* & Linda Riedemann Norbut**

The United States Supreme Court has failed to grapple with the unique interpretive difficulties presented by social media threats cases. Social media make hateful and threatening speech more common but also magnify the potential for a speaker's innocent words to be misunderstood. People speak differently on different social media platforms, and architectural features of platforms, such as character limits, affect the meaning of speech. The same is true of other contextual clues unique to social media, such as gifs, hashtags, and emojis. Only by understanding social media contexts can legal decision-makers avoid overcriminalization of speech protected by the First Amendment. This Article therefore advocates creation of a procedural mechanism for raising a "context" defense to a threats prosecution prior to trial. Comparable privileges protect defamation defendants from having their opinions misconstrued as defamatory and allow them to have their liability resolved at an early stage of litigation, often avoiding the anxiety and expense of trial. This Article contends that criminal defendants in threats cases should have a similar defense that permits them to produce contextual evidence relevant to the interpretation of alleged threats for consideration by a judge at a pretrial hearing. In cases that cannot be resolved before trial, the context defense would entitle a defendant to produce contextual evidence at trial and have the jury instructed regarding the role of context in separating threats from protected speech. Although adoption of the context defense would be especially helpful in correctly resolving social media cases, its use in all threats cases would provide an important safeguard against erroneous convictions of speech protected by the First Amendment.

DOI: https://doi.org/10.15779/Z38JM23G4C

* Lyrissa Lidsky is Dean and Judge C.A. Leedy Professor at the University of Missouri School of Law. She thanks Professor Michael Birnhack and the students at Tel Aviv University for critiquing an early draft of this Article. This Article also benefited from the comments of participants at Fordham

> [W]hatever it is that the law is after it is not the whole story.[1]
>
> --Clifford Geertz

INTRODUCTION

"*I think I'ma SHOOT UP A KINDERGARTEN/ AND WATCH THE BLOOD OF THE INNOCENT RAIN DOWN/ AND EAT THE BEATING HEART OF ONE OF THEM.*"[2] These were the words that Texas teen Justin Carter posted on Facebook in February 2013, just two months after a deranged gunman slaughtered twenty children and six adults at a Newtown, Connecticut school.[3] A Canadian onlooker was so alarmed that she took a screenshot of Carter's comments, looked up his address, and contacted Texas police.[4] Police officers responded promptly. Although they found no weapons of any kind in Carter's home, they arrested and charged him with making a terroristic threat, and a judge ultimately set bail at half a million dollars.[5] An anonymous donor posted bail,

Law Review's 2017 symposium on Terrorist Incitement on the Internet and the 2017 Yale Freedom of Expression Scholars Conference. Special thanks to Alexander Tsesis for his comments at both events and to Mitchell McNaylor for his early and excellent editorial advice. Thanks also to Ashley Messenger, Caroline Corbin, David Thaw, and Rachael Jones.

** Linda Riedemann Norbut is an attorney for the Brechner Center for Freedom of Information at the University of Florida, where she received her J.D. (2017) and M.A. in the Law of Mass Communication (2013).

1. CLIFFORD GEERTZ, LOCAL KNOWLEDGE: FURTHER ESSAYS IN INTERPRETIVE ANTHROPOLOGY 173 (1983).

2. Elise Hu, *As Supreme Court Considers Online Threats, an Update on Justin Carter*, NPR (Dec. 1, 2014), http://www.npr.org/sections/alltechconsidered/2014/12/01/367771533/as-high-court-considers-online-threats-an-update-on-justin-carterq [https://perma.cc/5XCX-KZPH].

3. *See* Joe Palazzolo, *Teen Jailed for Facebook Posting About School Shooting*, WALL ST. J. (July 4, 2013), https://www.wsj.com/articles/SB10001424127887324260204578583482948367730 [https://perma.cc/K9EY-UD4U].

4. *See* Hu, *supra* note 2.

5. *Id.*

but not before Carter had spent four months in jail, where he had been physically abused and put in solitary confinement for his own safety.[6] Carter awaited trial for five years before prosecutors finally gave him a plea deal in March 2018, dismissing the felony charges in exchange for his guilty plea to an unrelated misdemeanor charge.[7]

The justice system's response to Carter's words may sound predictable enough, even laudable, but setting Carter's words in their broader context suggests his real offense may have been speaking in bad taste and having bad timing.[8] The one comment from Carter's interlocutor included in the indictment suggests his "threat" may have been a hyperbolic response to provocation.[9] Moreover, Carter's father told interviewers that his son immediately followed his alleged threat with a post saying "'LOL' and 'J/K,'"[10] common internet abbreviations for "laughing out loud" and "just kidding."[11] Even Carter's use of selective capitalization can be read as a tip-off that his "threat" was merely a "rant," given that "[t]yping in all caps is Internet code for shouting."[12]

Other contextual evidence bolsters this interpretation. Although Carter made the offending comments on a public Facebook page, he did so in a war of

---

6. *Id.*

7. Austin Sanders, *Felony Charges Dropped in "Facebook Threat" Case*, AUSTIN CHRONICLE (Apr. 6, 2018), https://www.austinchronicle.com/daily/news/2018-04-06/felony-charges-dropped-in-facebook-threat-case/ [https://perma.cc/43YS-KYJM]. Carter pled guilty to a misdemeanor charge of filing a false report or alarm for which he was let off on time already served in jail. *Id.* Prior to the plea deal, the conditions of his bail prevented Carter from accessing the Internet (and thus from being employed) and prevented him from living with his underage siblings. *See* Hu, *supra* note 2. It is fair to conclude that his Facebook post, and what this Article will contend is a disproportionate response by the justice system, ruined his life. *See id.*

8. Katy Hollingsworth, *What Happened to Justin Carter, League of Legends 'Terrorist,'* GAME SKINNY (Feb. 13, 2014), http://www.gameskinny.com/ydr6b/what-happened-to-justin-carter-league-of-legends-terrorist [https://perma.cc/7V68-7SCJ].

9. Overall, the evidence suggests that Carter was moody and dark, often speaking of suicide. *Id.* While he was in jail, he was put on a suicide watch, and even prior to his unfortunate Facebook post, his high school girlfriend worried enough about his violent and suicidal musings that she sought a temporary restraining order against him. *Id.* Nonetheless, a search of his house turned up no weapons, nor any other indicia of a threat to carry out the "terroristic threats" with which he was charged. *Id.* Not that either the restraining order or Justin's inability to carry it out matter, technically, to the threats charge. *Id.* The restraining order is inadmissible character evidence, and a person can make a threat even if she never intends to carry it out. *See* FED. R. EVID. 404.

10. Brandon Griggs, *Teen Jailed for Facebook 'Joke' Is Released*, CNN (July 13, 2013), http://www.cnn.com/2013/07/12/tech/social-media/facebook-jailed-teen [https://perma.cc/ETW8-TTK5]. Neither police nor prosecutors produced the full series of posts in which Carter's comments appeared.

11. For a recent posting for a job as an emoji translator, see Alanna Petroff, *Now Hiring: Emoji Translator*, CNN (Dec. 13, 2016), http://money.cnn.com/2016/12/13/technology/emoji-translator-job-language [https://perma.cc/Z3SV-H6DH].

12. *See* Alice Robb, *How Capital Letters Became Internet Code for Yelling, and Why We Should Lay Off the All-Caps Key*, NEW REP. (Apr. 17, 2014), https://newrepublic.com/article/117390/netiquette-capitalization-how-caps-became-code-yelling [https://perma.cc/5VK5-5QJU].

words with a fellow player of League of Legends.[13] League is a multiplayer online battle game, played mostly by males between the ages of sixteen and thirty.[14] League players commonly engage in trash talk and hyperbolic exaggerations.[15] Indeed, it appears that Carter's interlocutor, also an "insider" of the gaming subculture, "trashed" him first, and although she was the immediate audience of the threat, she did not alert authorities. On the other hand, based on the fact that Carter's comment appeared on Facebook, a social media platform frequented by more middle-aged women than teenage boys,[16] some might argue that he should have been more circumspect before throwing around casual threats to kill kindergarteners. Regardless, establishing the full context of Carter's remarks should be an essential part of determining whether he made terroristic threats or merely talked trash with a fellow video gamer.

Unfortunately for Carter, the US Supreme Court's First Amendment jurisprudence has failed to resolve fundamental interpretive questions that should determine whether his words were a felony or free speech. In 1969, the Supreme Court declared, "What is a threat must be distinguished from what is constitutionally protected speech."[17] Yet, the Court has done relatively little to guide police, prosecutors, lower court judges, and juries seeking to make the necessary and difficult distinctions.[18] True, the Court has defined "true threats"

13. Alyson Shontell, *A Teen Was Jailed for a 'Sarcastic' Facebook Post Even Though the Cops Never Saw the Actual Conversation*, BUS. INSIDER (Feb. 14, 2014), http://www.businessinsider.com/justin-carters-facebook-comment-scandal-2014-2 [https://perma.cc/Q7FF-2QER].

14. League attracts over one hundred million active players per month. Ninety percent of League players are male, and 85 percent are between the ages of sixteen and thirty. Drew Harwell, *More Women Play Video Games than Boys, and Other Surprising Facts Lost in the Mess of Gamergate*, WASH. POST (Oct. 17, 2014), https://www.washingtonpost.com/news/the-switch/wp/2014/10/17/more-women-play-video-games-than-boys-and-other-surprising-facts-lost-in-the-mess-of-gamergate [https://perma.cc/LG85-TVHJ]; Anthony Gallegos, *Riot Games Releases Awesome League of Legends Infographic*, IGN (Oct. 15, 2012), http://www.ign.com/articles/2012/10/15/riot-games-releases-awesome-league-of-legends-infographic [http://perma.cc/BU5D-CH4R].

15. Trash talk is common in other contexts. For example, David W. Rainey and Vincent Granito examined trash talk in college athletics and found that "there are normative rules favoring trash talk among collegiate athletes and that the rules vary somewhat by gender, level of competition, and sport." David W. Rainey & Vincent Granito, *Normative Rules for Trash Talk Among College Athletes: An Exploratory Study*, 33 J. SPORT BEHAV. 276, 276 (2010).

16. According to a Pew Internet poll, while young adults continue to report using Facebook at high rates, older adults are joining in increasing numbers. In addition, women continue to use Facebook at higher rates than men. Shannon Greenwood, Andrew Perrin & Maeve Duggan, *Social Media Update 2016*, PEW RES. CTR. 4 (Nov. 11, 2016), http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/11/10132827/PI_2016.11.11_Social-Media-Update_FINAL.pdf [https://perma.cc/9W6F-9V8F].

17. Watts v. United States, 394 U.S. 705, 707 (1969).

18. *See, e.g.*, Alec Walen, *Criminalizing Statements of Terrorist Intent: How to Understand the Law Governing Terrorist Threats, and Why It Should Be Used Instead of Long-Term Preventive Detention*, 101 J. CRIM. L. & CRIMINOLOGY 803, 825 (2011) (contending that "the current doctrine of true threats is incoherent"); Clay Calvert & Matthew D. Bunker, *Fissures, Fractures & Doctrinal Drifts: Paying the Price in First Amendment Jurisprudence for a Half Decade of Avoidance, Minimalism & Partisanship*, 24 WM. & MARY BILL OF RTS. J. 943, 957, 959 (2016) (noting the Supreme Court's failure

as "statements where the speaker means to communicate a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals."[19] Further, the Court has explained that the First Amendment allows the criminalization of threatening speech because it causes fear, social disruption, and heightens the risk of future violence.[20] The Court has failed, however, to answer fundamental questions regarding the "true threats exception" to First Amendment protection, including whether courts should view threats from the vantage of the speaker, a reasonable recipient, a reasonable disinterested reader, or all of the above;[21] and what mens rea the First Amendment requires in

---

to clarify the true threats doctrine in *Elonis v. United States*, 135 S. Ct. 2001 (2015), and stating that "[i]f one First Amendment doctrine screams out the loudest for clarification, it may well be true threats.").

19. Virginia v. Black, 538 U.S. 343, 359 (2003).

20. *Id.* at 360. The Court also justified criminalization of threats on the grounds of protecting people "from the possibility that the threatened violence will occur." *Id*. (quoting R.A.V. v. City of St. Paul, 505 U.S. 377, 388 (1992)).

21. In the absence of adequate guidance from the Supreme Court, lower courts have adopted various tests for judging whether a statement is a threat. Some have based their tests on the subjective perspective of the speaker, some on the objectively reasonable speaker, some on the reasonable recipient, and some on a combination of subjective and objective perspectives. Prior to the Supreme Court's decision in *Elonis*, a majority of circuit courts required the state to prove that the defendants intended to make a statement that could reasonably (objectively) be viewed as a threat, either judged from the perspective of a reasonable speaker, a reasonable recipient, or both. See *United States v. Martinez*, 736 F.3d 981, 985–86, 988 (11th Cir. 2013), applying:

> an objective, reasonable-person test when distinguishing true threats from protected speech. Under that objective standard, a true threat is a communication that, when taken in context, 'would have a reasonable tendency to create apprehension that its originator will act according to its tenor.' . . . Knowingly transmitting the threat makes the act criminal—not the specific intent to carry it out or the specific intent to cause fear in another.

Id. (first citing United States v. Callahan, 702 F.2d 964, 965 (11th Cir. 1983); then quoting United States v. Alaboud, 347 F.3d 1293, 1296–97 (11th Cir. 2003); and then quoting United States v. Fuller, 387 F.3d 643, 646 (7th Cir. 2004)); United States v. Fuller, 387 F.3d 643, 646 (7th Cir. 2004); United States v. Elonis, 730 F.3d 321, 328 (3d Cir. 2013) (rejecting "a subjective intent requirement that the defendant 'intended at least to convey the impression that the threat was a serious one,'" and holding that a true threat requires the speaker to "knowingly and willfully" "make a statement, written or oral, in a context or under such circumstances wherein a reasonable person would foresee that the statement would be interpreted by those to whom the maker communicates the statement as a serious expression of an intention to inflict bodily harm") (quoting United States v. Kosma, 951 F.2d 549, 557–58 (3d Cir. 1991)); United States v. Mabie, 663 F.3d 322, 330, 332 (8th Cir. 2011) ("A 'true threat' is defined as a 'statement that a reasonable recipient would have interpreted as a serious expression of an intent to harm or cause injury to another.'" This objective test "does not consider the subjective intent of the speaker.") (quoting Doe v. Pulaski Cty. Special Sch. Dist., 306 F.3d 616, 624 (8th Cir. 2002)); United States v. Stewart, 411 F.3d 825, 828 (7th Cir. 2005) ("Whether the letter contains a 'true threat' is an objective inquiry. In other words, guilt is not dependent upon 'what the defendant intended, but whether the recipient could reasonably have regarded the defendant's statement as a threat.'") (quoting United States v. Aman, 31 F.3d 550, 553 (7th Cir. 1994)); Porter v. Ascension Parish Sch. Bd., 393 F.3d 608, 616 (5th Cir. 2004) ("Speech is a 'true threat' and therefore unprotected if an objectively reasonable person would interpret the speech as a 'serious expression of an intent to cause a present or future harm.'") (quoting *Doe*, 306 F.3d at 622); *Alaboud*, 347 F.3d at 1297 & n.3 (A statement is a true threat if "a reasonable person [would] construe it as a serious intention to inflict bodily harm."); United States v. Jeffries, 692 F.3d 473, 479 (6th Cir. 2012) ("[A] threat must be communicated with intent (defined objectively) to intimidate."). Other courts required that the state prove that the defendant subjectively intended her statement to be a threat. *See* United States v. Cassel, 408 F.3d 622, 631 (9th Cir. 2005) (A true threat

threats cases.[22] The inadequacy of current true threats doctrine is especially acute in the social media era.[23] As billions of people have flocked to social media,[24] the amount of threatening and hateful speech to be found there has grown "massively."[25] Meanwhile, as Justin Carter's story illustrates, the Court's failure to clarify its true threats doctrine has real consequences for real people.

This Article seeks to remedy some of the doctrinal deficits highlighted by the growth of incendiary speech in social media.[26] Along the way, this Article

---

"requires that 'the speaker means to communicate . . . an intent to commit an act of unlawful violence.'" This means that "the communication itself [must] be intentional," and "that the speaker intended for his language to threaten the victim.") (emphasis omitted); United States v. Fulmer, 108 F.3d 1486, 1491 (1st Cir. 1997) (A true threat depends on "whether [the speaker] should have reasonably foreseen that the statement he uttered would be taken as a threat by those to whom it is made.").

22. *See* Paul T. Crane, Note, *"True Threats" and the Issue of Intent*, 92 VA. L. REV. 1225, 1227–29 (2006) (noting that these matters are unresolved, as well as the question whether threats must be specific or imminent in order to be actionable). Leading Cases, *Federal Threats Statute—Mens Rea and the First Amendment*—Elonis v. United States, 129 HARV. L. REV. 331 (2015) (noting that it is unresolved whether the First Amendment requires a defendant to have subjectively intended to threaten someone).

23. *See, e.g.*, Lyrissa Barnett Lidsky, *Incendiary Speech and Social Media*, 44 TEX. TECH. L. REV. 147, 147–49 (2011) (noting that "a number of factors potentially contribute to the incendiary capacity of social media speech," and "underlying assumptions" regarding how "audiences respond to incitement, threats, or fighting words, are confounded by the reality social media create"); Eric J. Segall, *The Internet as a Game Changer: Reevaluating the True Threats Doctrine*, 44 TEX. TECH. L. REV. 183, 184 (2011) ("[T]he Internet is a game changer when it comes to criminal law and free speech [because] there is simply no pre-Internet analogy that allows speech to be disseminated so quickly, so cheaply, and to so many for such a long period of time."); TIMOTHY GARTON ASH, FREE SPEECH: TEN PRINCIPLES FOR A CONNECTED WORLD 220 (2016) ("The internet has brought an explosion of offensive, extreme expression, exacerbated by the online norm of anonymity.").

24. Dave Chaffey, *Global Social Media Statistics Summary 2018*, SMART INSIGHTS (Mar. 28, 2018), http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research [https://perma.cc/MJ9K-8G8R]; *Number of Social Media Users Worldwide 2010–2020*, STATISTA, https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users [https://perma.cc/R9QX-JLFV]. As Professor Jennifer E. Rothman observes, "interest in threats" has sharpened since the advent and "proliferation of widely disseminated Internet speech." Jennifer E. Rothman, *Freedom of Speech and "True Threats*,*"* 25 HARV. J.L. & PUB. POL'Y 283, 286 (2001). Some attribute the increasing amount of hateful speech to growing political polarization. *Political Polarization*, PEW RES. CTR. RSS (2016), http://www.pewresearch.org/topics/political-polarization [https://perma.cc/Z9MP-3RP9].

25. Caitlin Dickerson, *Reports of Bias-Based Attacks Tick Upward After Election*, N.Y. TIMES (Nov. 11, 2016), http://www.nytimes.com/2016/11/12/us/reports-of-bias-based-attacks-tick-upward-after-election.html [https://perma.cc/KUP5-PDQW]; Jessica Guynn, *'Massive Rise' in Hate Speech on Twitter During Presidential Election*, USA TODAY (Oct. 23, 2016), http://www.usatoday.com/story/tech/news/2016/10/21/massive-rise-in-hate-speech-twitter-during-presidential-election-donald-trump/92486210 [https://perma.cc/8QVC-V2HX]. The rising rates of hate speech are not limited to the United States. Scott Roxborough, *Online Hate Speech Soars in Germany*, HOLLYWOOD REP. (Nov. 28, 2016), http://www.hollywoodreporter.com/news/online-hate-speech-soars-germany-950657 [https://perma.cc/5J7L-8UUC]; Mike Wendling, *2015: The Year that Angry Won the Internet*, BBC TRENDING (Dec. 30, 2015), http://www.bbc.com/news/blogs-trending-35111707 [https://perma.cc/J6CU-W48Q].

26. *See* Alexander Tsesis, *Inflammatory Speech: Offense Versus Incitement*, 97 MINN. L. REV. 1145, 1173 (2013) ("The applicability of the true threats doctrine to Internet communication has been woefully understudied.").

borrows insights from defamation and other tort cases, emphasizing the importance of context in separating protected and unprotected speech, and the necessity of independent review of jury determinations of both context and intent.

Part I examines the Supreme Court's limited body of "true threats" jurisprudence. This analysis demonstrates that the lens through which the Court has viewed context in its threats jurisprudence assumes that the Justices, the immediate audience of the purported threat, and anyone interpreting the words at the time they were said—or the time the jury interpreted them—shared a common social, cultural, linguistic, and political frame of reference. Even the Court's recent Facebook threats case, *Elonis v. United States*,[27] failed to explain how to interpret threatening speech in situations in which legal decision-makers do not share a frame of linguistic reference with the speaker or her audience. This deficit, coupled with the doctrinal uncertainty regarding the requisite mens rea, leaves legal decision-makers at a loss in separating social media threats from violent, yet protected, free expression.

Part II explains why context matters when courts evaluate threats made in social media. This Section explores the facets of social media that make dangerous and destructive speech appear to be more common as well as the factors that make it harder to discern threatening speech from hyperbole. The spontaneous, informal, unmediated, and often-anonymous nature of social media gives users license to say things online that they would never say in person and contributes to the harm suffered by targets of both hateful speech and true threats. Yet, these same characteristics magnify the potential for a speaker's innocent words to be misunderstood. Misunderstandings may also arise because traditional context clues signaling a speaker's intent are absent, replaced by new clues that may be difficult to decode, such as hashtags, emojis, and gifs. Moreover, different social media platforms have different discourse conventions and architectural features which complicate attempts to discern both the speaker's true intent and the meaning of her postings. Finally, speakers of different ages and backgrounds use social media differently to convey their messages, adding another layer of contextual complexity.

Part III considers objections to requiring contextual analysis of social media "threats." Many would contend that the First Amendment need not extend "breathing space" to posts like Justin Carter's: even if Carter did not intend for his threat to be taken seriously and his interlocutor did not in fact take his threat seriously, his comments were still capable of causing social disruption when taken out of context. Some would even argue that being misconstrued is a risk that speakers like Carter should have to assume. Part III, however, asserts that the First Amendment demands precise line-drawing even for threatening speech like Carter's. Further, this Section demonstrates that it is dangerous, from a First

27. 135 S. Ct. 2001 (2015).

Amendment perspective, to give police and prosecutors a broad mandate to punish fear-inducing speech by speakers from sub-communities perceived as deviant. As a result, threatening speech should only be actionable if the speaker intends her words to threaten or intimidate,[28] and her intended targets would reasonably perceive that intent. In analyzing both intent and effect, context matters.

Therefore, in order to help legal decision-makers face the challenges attendant to interpreting social media threats, Part IV first identifies contextual analysis guideposts that should anchor the analysis of true threats. Then, it proposes the creation of a procedural mechanism for raising a "context" defense to the prosecution of a threat-based offense prior to trial. Comparable privileges protect defamation defendants from having opinions misconstrued as defamation and allow them to have their liability resolved at an early stage, often before they must undergo the anxiety and expense of trial. This Article, therefore, proposes a new defense in threats cases: one that would allow them to produce contextual evidence relevant to the interpretation of alleged threats for consideration by a judge at a pretrial hearing. In cases where contextual issues cannot be resolved pretrial, the proposed context defense would entitle the defendant to produce evidence of context at trial and have the jury be instructed regarding the critical role of context in separating threats from protected speech.[29]

I.
WATTS AND ITS PROGENY

From the very inception of the true threats doctrine in *United States v. Watts*,[30] the Supreme Court has paid insufficient attention to how context affects the interpretation of allegedly threatening speech. Perhaps this oversight stems in part from the fact that *Watts* was a relatively easy case.

Like Justin Carter, the defendant in *Watts* was an eighteen-year-old whose rash words landed him in trouble.[31] Robert Watts made his allegedly threatening comments at a 1966 political rally at the Washington Monument.[32] At the time, opposition to US involvement in the Vietnam War was widespread, and vituperative attacks on the President were common.[33] For example, a *New York Times* article from February 1966 recounts that four thousand people protested

28. Adrienne Scheffey, *Defining Intent in 165 Characters or Less*, 69 U. MIAMI L. REV 861, 893 (2015) ("Particularly when assessing online speech, anonymous or not, difficulties arise in applying the reasonable person test.").

29. As detailed in Part IV *infra*, the prosecution would maintain the burden of proving the requisite intent. In other words, this privilege acts as a "bursting bubble" presumption. *See* Mason Ladd, *Presumptions in Civil Actions*, 1977 ARIZ. ST. L.J. 281–82 (1977) ("Under the Thayer-Wigmore rule, commonly designated as the 'bursting bubble' theory, the presumption vanishes when evidence is introduced which would support a finding of the nonexistence of the presumed fact.").

30. 394 U.S. 705 (1969).

31. *Id.* at 705–06.

32. *Id.*

33. *Id.*

outside the hotel where then-President Lyndon Baines Johnson was speaking, chanting, "Hey, hey, LBJ, how many kids did you kill today?"[34]

It was in this larger political context that Watts declaimed to a small group of listeners, later described by the US Supreme Court as mostly in their teens or early twenties:

> They always holler at us to get an education. And now I have already received my draft classification as 1-A[35] and I have got to report for my physical this Monday coming. I am not going. If they ever make me carry a rifle the first man I want to get in my sights is L.B.J.[36]

The audience laughed, but an investigator for the Army Counter Intelligence Corps attending the rally was not amused.[37] A District of Columbia jury convicted Watts of threatening the life of the President.[38] Watts's lawyer sought a judgment of acquittal on the basis that the comments could not be interpreted as a threat.[39] Neither the trial judge nor jury accepted this argument.[40] Luckily for Watts, a majority of Justices on the Supreme Court did.[41]

In a per curiam opinion, the Court relied on its own understanding of the surrounding political, social, and linguistic context of Watts's speech to second-guess the jury that convicted him and labeled his comments as mere "political hyperbole."[42] The Court readily viewed Watts's expression as a valuable

---

34. Martin Arnold, *4,000 Picket Johnson in Antiwar Protest at Hotel*, N.Y. TIMES (Feb. 24, 1966), https://www.nytimes.com/1966/02/24/archives/4000-picket-johnson-in-antiwar-protest-at-hotel-they-chant-peace.html [https://perma.cc/RV3K-U2SU].

35. A 1-A classification indicates a person is available for military service. *See Classifications*, SELECTIVE SERV. SYS., https://www.sss.gov/Classifications [https://perma.cc/CGM6-HQ95].

36. *Watts*, 394 U.S. at 706.

37. *Id.*

38. *Id.* The statute that Watts was convicted of violating, 18 U. S. C. § 871(a), provides in full:

> Whoever knowingly and willfully deposits for conveyance in the mail or for a delivery from any post office or by any letter carrier any letter, paper, writing, print, missive, or document containing any threat to take the life of or to inflict bodily harm upon the President of the United States, the President-elect, the Vice President or other officer next in the order of succession to the office of President of the United States, or the Vice President-elect, or knowingly and willfully otherwise makes any such threat against the President, President-elect, Vice President or other officer next in the order of succession to the office of President, or Vice President-elect, shall be fined under this title or imprisoned not more than five years, or both.

18 U. S. C. § 871(a) (2012).

39. *Watts*, 394 U.S. at 706–07 ("He stressed the fact that petitioner's statement was made during a political debate, that it was expressly made conditional upon an event-induction into the Armed Forces-which petitioner vowed would never occur, and that both petitioner and the crowd laughed after the statement was made.").

40. *Id.* at 705–06.

41. *Id.* at 707.

42. *Id.* at 708. By the time the Court decided the case in 1969, President Johnson had declined to seek reelection, largely because he knew he could not win due to his role in prosecuting the Vietnam War. *See* Robert Mitchell, *A 'Pearl Harbor in Politics': LBJ's Stunning Decision Not to Seek Reelection*, WASH. POST (Mar. 31, 2018), https://www.washingtonpost.com/news/retropolis/wp/2018/03/31/a-

contribution to political debate rather than a true threat,[43] opining "we do not see how it could be interpreted otherwise."[44] The Court bolstered its conclusion by a dissection of the surrounding context. As the Court observed, Watts made his statements in front of a small group of young adults engaged in political discussion.[45] The Court also explicitly referenced the audience's reaction to Watts's statement—they laughed—and the precise wording of the statement was "expressly conditional."[46] Thus, the Court concluded that Watts's "only offense here was 'a kind of very crude offensive method of stating a political opposition to the President.'"[47] In interpreting Watts's statement as mere hyperbole, the Court also referenced a significant number of contextual clues, including his tone and sentence structure, the physical setting, the age and emotional characteristics of the immediate audience, the relationships between Watts and his immediate audience, and perhaps most importantly, the broader social and political context.[48] *Watts*, however, was an easy case, and the Court viewed it as such, in part because the Court's opinion shows that it readily understood the conventions of late 1960s-era political discourse by young people. The fact that the Court viewed *Watts* as an easy case[49] also perhaps explains why it sets forth no elaborate definitions of true threats, rationales for excluding true threats from First Amendment protection, or multi-factor tests for distinguishing threats from protected speech.[50] Unfortunately, the Court waited thirty-four more years after

pearl-harbor-in-politics-lbjs-stunning-decision-not-to-seek-reelection [https://perma.cc/R7CE-WT8X] ("Johnson, however, saw . . . . [that] he was not likely to win . . . . because of the Vietnam War.").

43. The Court cited *New York Times v. Sullivan* for the proposition that our "profound national commitment to the principle that debate on public issues should be uninhibited, robust, and wide-open, and that it may well include vehement, caustic, and sometimes unpleasantly sharp attacks on government and public officials." *Watts*, 394 U.S. at 708 (quoting N. Y. Times v. Sullivan, 376 U.S. 254, 270 (1964)). The Court further acknowledged that in "the political arena" or "labor disputes," "vituperative, abusive, and inexact" language is to be expected. *Id.* (citing Linn v. United Plant Guard Workers of Am., 383 U.S. 53, 58 (1966)).

44. *Id.* at 708.

45. *Id.*

46. *Id.*

47. *Id.*

48. *Id.*

49. *Hess v. Indiana* was another relatively easy case because the Court readily understood the surrounding context. *See* 414 U.S. 105, 107 (1973) (per curiam). There, the Court struck down a disorderly conduct conviction of an anti-war demonstrator for saying, "We'll take the fucking street later." *Id.* As in *Watts*, the Court looked at the surrounding context and determined that Hess's statement was not directed to any individual or group and that "there was no evidence or rational inference from the import of the language, that his words were intended to produce, and likely to produce, imminent disorder." *Id*. at 108–09 (emphasis omitted). The *Hess* case involved not threats but incitement, as defined by *Brandenberg v. Ohio*, 395 U.S. 444 (1969), but the Court's decision nonetheless illustrates its willingness to cull contextual clues from the record to guide its interpretation of speech. *See id.* at 108.

50. *See* Tsesis, *supra* note 26, at 1161 ("Given that *Watts* overturned the conviction for threatening the president but confirmed the constitutionality of a statute that criminalized intentional intimidation, the Court spawned obscurity about what constituted a true threat.").

Watts before deciding some of these questions in a second threats case, and even when it did so, it left fundamental questions unresolved.

Virginia v. Black[51] remains the Court's most definitive statement of the true threats doctrine and most complete explanation of why the First Amendment permits the government to criminalize true threats.[52] In *Black*, the Court finally defined "true threats" as statements that manifest a speaker's "serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals."[53] The Court also explained why criminalizing threats is constitutionally permissible.[54] Criminalizing threats "'protect[s] individuals from the fear of violence' and 'from the disruption that fear engenders,' in addition to protecting people 'from the possibility that the threatened violence will occur.'"[55] Despite this seeming attempt to clarify the true threats doctrine, the ambiguity of the proffered definition, together with the unusual facts of *Black*, sowed significant confusion and ultimately produced no judicial consensus regarding how to draw the line between threats and protected speech.

Virginia v. Black involved the constitutionality of a state statute singling out a particular "subset" of threats for criminalization, namely cross burning "with the intent of intimidating any person or group of persons."[56] The statute made cross burning "prima facie evidence of an intent to intimidate."[57] *Black* involved several respondents, one of whom was a Ku Klux Klansman convicted of violating the statute after his jury was instructed that his burning of a cross was sufficient evidence from which they might infer a motive "to intentionally put a person or group of persons in fear of bodily harm."[58]

---

51. 538 U.S. 343 (2003).

52. *Id.* at 360.

53. *Id.* at 359.

54. Actually, the Court reiterated dicta from the "hate speech" case of *R.A.V. v. City of St. Paul*, 505 U.S. 377, 388 (1992), on this issue. *See Black*, 538 U.S. at 360.

55. *Id.* at 360 (citing *R.A.V.*, 505 U.S. at 388). These rationales, while useful touchstones, surely do not permit criminalization of all speech that causes fear or social disruption or even speech that makes violence more likely. *See, e.g.*, Brandenburg v. Ohio, 395 U.S. 444 (1969).

56. The statute provided:

> It shall be unlawful for any person or persons, with the intent of intimidating any person or group of persons, to burn, or cause to be burned, a cross on the property of another, a highway or other public place. Any person who shall violate any provision of this section shall be guilty of a Class 6 felony. Any such burning of a cross shall be prima facie evidence of an intent to intimidate a person or group of persons.

VA. CODE ANN. § 18.2-423 (1996).

57. *Black*, 538 U.S. at 348.

58. *Id*. at 349. The other two defendants challenging their convictions before the Court apparently sought to intimidate an African-American neighbor by burning a cross on his lawn. *Id.* at 350. A Virginia jury found one of them guilty of attempted cross burning under the Virginia statute. *Id.* at 351. The other pleaded guilty but preserved the right to challenge the constitutionality of the cross-burning statute. *Id.* at 350.

A majority of Supreme Court justices held that states may criminalize cross burning done with the intent to intimidate,[59] but a plurality held that the cross-burning statute at issue violated the First Amendment by permitting juries to *presume* intent merely from the act of cross burning itself.[60] The presumption of intent was a procedural flaw[61] that allowed juries to "ignore[] *all of the contextual factors that are necessary* to decide whether a particular cross burning is *intended* to intimidate. The First Amendment does not permit such a shortcut."[62] In other words, the Court held that a state may punish cross burning done with the intent to intimidate but may not punish cross burning done for purely ideological or artistic reasons.[63]

The Court held that even though the First Amendment normally forbids content-based regulation of expression,[64] states may engage in content-based regulation targeting those threats "most likely to inspire fear of bodily harms."[65] Justice Sandra Day O'Connor, writing for the Court, squarely placed cross burning with the intent to intimidate in this category, describing it as a "type of true threat, where a speaker directs a threat to a person or group of persons *with the intent of placing the victim in fear of bodily harm or death*."[66] To bolster this conclusion, she gave a detailed history of the practice of cross burning, [67] its cultural uses by the Klan as a tool of intimidation, and its frequent link to violence.[68] She also described the effect of cross burning on victims (or perhaps on observers): "cross burning is often intimidating, intended to create a pervasive

59. *Id.* at 363. Justice O'Connor was joined in this part of the opinion by Chief Justice Rehnquist and Justices Stevens, Scalia, and Breyer. *Id.* at 347. Justice Thomas concurred in the result as well. *Id.* Justice Souter, joined by Justices Kennedy and Ginsburg, dissented on the grounds that no exception, including such a rule that allows the prohibition of "particularly virulent" proscribable expression, could save the statute from unconstitutionality. *Id.* at 382 (Souter, J., dissenting).

60. *Id.* at 367. The plurality, comprised of Justice O'Connor, Chief Justice Rehnquist, and Justices Stevens and Breyer, concluded that the prima facie evidence provision in the statute made it unconstitutionally overbroad. *Id.* at 365.

61. Timothy Zick, *Cross Burning, Cockfighting, and Symbolic Meaning: Toward a First Amendment Ethnography*, 45 WM. & MARY L. REV. 2261, 2264 (2004) ("Although the Court concluded that the government could prohibit cross burning as a form of threatening symbolism, it held that the Virginia statute was procedurally flawed.").

62. *Black*, 538 U.S. at 367 (emphasis added).

63. *See id.* at 398 (Thomas, J., dissenting) (questioning whether the plurality's invalidation of the prima facie evidence provision showed undue concern for the "innocent cross-burner").

64. *See* R.A.V. v. City of St. Paul, 505 U.S. 377, 382 (1992).

65. *Black*, 538 U.S. at 363. *But see R.A.V.*, 505 U.S. at 393–94 (determining that content-based regulations within an unprotected category of speech, such as fighting words, are impermissible if the state aims to proscribe the mere expression of particular ideas without curbing "a particularly intolerable (and socially unnecessary) *mode* of" expression).

66. *Black*, 538 U.S. at 359–60 (emphasis added).

67. *Id.* at 352–57. Professor Guy-Uriel Charles explains that the Court's historical overview in *Black* illustrates several important lessons, including the following: "First, cross burning in the United States is uniquely associated with the Ku Klux Klan. Second, cross burning is a harbinger of impending violence and is usually followed by violence. Third, a critical purpose of cross burning is intimidation." Guy-Uriel E. Charles, *Colored Speech: Cross Burnings, Epistemics, and the Triumph of the Crits?*, 93 GEO. L.J. 575, 628–29 (2005) (arguing *Black* cannot be reconciled with *R.A.V.*).

68. *Black*, 538 U.S. at 352–55.

fear in victims that they are a target of violence."[69] The Court recognized that cross burning had other potential meanings; for example, for those within the Klan, cross burning operates as a "message[] of shared ideology."[70] Despite this, states can still punish cross burning because it is a kind of "embodied threat" that is especially intimidating.[71] States may not, however, punish cross burning based solely on antipathy to any ideological message conveyed.[72]

The Court's decision in *Black* might be read as an endorsement of a broad contextual analysis of expressive conduct alleged to be a threat. However, the Court's insistence on cross burning as a historically and culturally exceptional kind of threatening conduct limits the decision's applicability to other threats cases. Moreover, the Court left unresolved even more elemental questions about true threats: from whose vantage point threats should be judged, and relatedly, whether the test for threats is objective or subjective. *Black* provides uncertain answers to these questions, in part because the Virginia statute at issue in the case did not directly address threats writ large but instead addressed only certain types of cross burnings—those done with intent to intimidate.

Black implies that a defendant's subjective intent is an important part of the dividing line between threatening and non-threatening cross burning, but it fails to address whether it is an important part of the dividing line in all threats cases.[73] The Court's definition of threats exacerbates the confusion because the definition can be read either as prioritizing the subjective perspective of the speaker or the objectively reasonable recipient in judging what is a threat.

Looking back at the definition, threats are statements that manifest a speaker's "serious expression of an intent to commit an act of unlawful violence

69. *Id*. at 360. Justice Thomas's powerful dissent suggests that the correct vantage point may be those who have historically been the targets of cross burnings. *See id.* at 389 (Thomas, J., dissenting) ("[T]he majority's brief history of the Ku Klux Klan only reinforces th[e] common understanding of the Klan as a terrorist organization, which, in its endeavor to intimidate, or even eliminate those it dislikes, uses the most brutal of methods.").

70. *Id.* at 354.

71. *Id.* at 355. *See* Alexander Tsesis, *Burning Crosses on Campus: University Hate Speech Codes*, 43 CONN. L. REV. 617, 630 (2010) ("True threats fall outside the[] accepted bounds of self-assertion because they are meant to menace someone with physical harm.").

72. The Court has similarly held that states may not ban flag burning based on antipathy to the message conveyed by the flag burner, but states may do so under content-neutral ordinances. *See* Texas v. Johnson, 491 U.S. 397, 412 & n.8 (1989) (explaining that an arson statute, for example, might have permissibly restricted flag burning).

73. For commentators criticizing the lack of clarity in the Court's definition of true threats, see Roger C. Hartley, *Cross Burning—Hate Speech as Free Speech: A Comment on* Virginia v. Black, 54 CATH. U. L. REV. 1, 2 (2004); W. Wat Hopkins, *Cross Burning Revisited: What the Supreme Court Should Have Done in* Virginia v. Black *and Why It Didn't*, 26 HASTINGS COMM. & ENT. L.J. 269, 308–09 (2004); Frederick Schauer, *Intentions, Conventions, and the First Amendment: The Case of Cross-Burning*, 2003 SUP. CT. REV. 197, 216 (2003); James L. Swanson, *Unholy Fire: Cross Burning, Symbolic Speech, and the First Amendment* Virginia v. Black, 2003 CATO SUP. CT. REV. 81, 95 (2003). Other commentators criticized the case for its failure to clarify the line between incitement and threats. *See, e.g.*, Steven G. Gey, *A Few Questions About Cross Burning, Intimidation, and Free Speech*, 80 NOTRE DAME L. REV. 1287, 1325–31 (2005).

to a particular individual or group of individuals."[74] If one focuses on threats as "*statements that manifest* a speaker's serious expression" of intent, then an objective "reasonable speaker" or "reasonable recipient" test for judging threats seems appropriate. Indeed, a majority of lower federal courts adopted such a test in the wake of *Virginia v. Black*.[75] On the other hand, if one focuses on the fact that the statements must manifest or contain "a speaker's *serious expression of an intent*," the focus shifts to the subjective intent of the speaker to judge the threat. Some lower courts have indeed read *Black* to require subjective intent on the part of the speaker, and some have read *Black* to require both that a speaker have the subjective intent to threaten and that the statement is in fact threatening, judged from the perspective of the reasonable speaker or recipient.[76] In other words, *Black* resulted in a circuit split regarding what type of intent a speaker must have in order to be held liable for his threatening speech.

Many observers believed the Supreme Court would repair this split when it granted certiorari in its first social media threats case in 2014.[77] When the Court granted certiorari in *Elonis v. United States*, it appeared the Justices would finally decide key questions about threats and intent, including whether the defendant must have general intent to communicate the threat or specific intent to threaten or intimidate, and if the latter, whether the requisite intent for constitutional purposes should be purpose, knowledge, recklessness, or negligence. The Court also seemed as if it might address from whose vantage point threats should be judged as well as whether social media threats might be different from their offline counterparts.[78] Unfortunately, the Supreme Court did none of these.

Elonis arose from a series of alleged threats made on Facebook by an angry, divorcing husband.[79] When Anthony Douglas Elonis's wife left him and took

74. *Black*, 538 U.S. at 359.

75. *See Recent Cases—First Amendment*, 126 HARV. L. REV. 1138, 1138 (2013) (discussing conflicting standards applied by lower courts in true-threat cases); Jake Romney, Note, *Eliminating the Subjective Intent Requirement for True Threats in* United States v. Bagdasarian, 2012 BYU L. REV. 639, 639 (2012) (same). *See also supra* note 21 and accompanying text.

76. *See supra* note 21.

77. Elonis v. United States, 135 S. Ct. 2001, 2011–12 (2015). The question was whether 18 U.S.C. § 875(c) (2012), which makes it a crime to transmit in interstate commerce "any communication containing any threat . . . to injure the person of another," "requires that the defendant be aware of the threatening nature of the communication, and—if not—whether the First Amendment requires such a showing." *Id.* at 2004. Clay Calvert and Matthew Bunker have observed that *Elonis* "provided the Court with a prime opportunity to resolve the circuit split on the question of intent, as well as to address whether (and how) posting messages to online social media platforms, such as Facebook, affects the true threats analysis." Calvert & Bunker, *supra* note 18, at 959; *see also* Christine LiCalzi, *Computer Crimes*, 54 AM. CRIM. L. REV 1025, 1039 (2017) (noting that the *Elonis* Court "declined to resolve the circuit split by deciding the case on statutory grounds").

78. *See* P. Brooks Fuller, *Evaluating Intent in True Threats Cases: The Importance of Context in Analyzing Threatening Internet Messages*, 37 HASTINGS COMM. & ENT. L.J. 37, 53 (2015) ("Ideally, the Court should address in its discussion, if not in its holding, whether an online communication is itself a contextually relevant factor and whether usage of social media impacts the intended or objective meaning of a communication.").

79. *Elonis*, 135 S. Ct. at 2004.

their kids, he vented his rage and frustration on Facebook.[80] He posted violent and disturbing words and images that earned him a federal indictment for threatening his wife, an FBI agent, a kindergarten, and patrons and co-workers at the park where he worked.[81] Unlike the Justin Carter case, Elonis's conduct involved *repeated* threats against a number of *specifically named* targets.[82] One count of the indictment against Elonis, for example, was based on a post in which he wrote, in "rap-style" lyrics, that the restraining order his wife obtained against him based on his prior Facebook posts would not protect her.[83] "Is it thick enough to stop a bullet?," he queried.[84] He asserted that the restraining order was "improperly granted" because "the Judge needs an education on true threats jurisprudence," hinting that he could win a settlement because his speech had been unconstitutionally suppressed.[85] He followed, however, by saying:

> And if worse comes to worse
> I've got enough explosives
> to take care of the State Police and the Sheriff's Department.[86]

Below his post, he linked to a Wikipedia entry on free speech,[87] but this link did little to negate the overall impression created by the violent and almost unhinged rhetoric in the post.

Another of Elonis's threatening posts presented a disturbing juxtaposition of taunts directed at law enforcement authorities, extremely violent imagery targeted at a specific woman, invocations of free speech rights, half-hearted disclaimers, and a seeming promise not to be deterred from violence.[88] That post came after Elonis received a visit at his home from FBI agents because he had posted about shooting up a school.[89] During the visit, Elonis was evidently "polite but uncooperative."[90] Shortly afterward, however, Elonis titled a Facebook post "Little Agent Lady," which the Supreme Court quoted as follows:

> "You know your s***'s ridiculous
> when you have the FBI knockin' at yo' door
> Little Agent lady stood so close
> Took all the strength I had not to turn the b**** ghost
> Pull my knife, flick my wrist, and slit her throat
> Leave her bleedin' from her jugular in the arms of her partner

80. *Id.* at 2006.
81. *Id.* at 2005.
82. *Id.* at 2005–07.
83. *Id.* at 2005.
84. *Id.* at 2006.
85. *Id.*
86. *Id.* (citing the indictment).
87. *Id.*
88. *Id.* at 2006–07.
89. *Id.*
90. *Id.*

> [laughter]
> So the next time you knock, you best be serving a warrant
> And bring yo' SWAT and an explosives expert while you're at it
> Cause little did y'all know, I was strapped wit' a bomb
> Why do you think it took me so long to get dressed with no shoes on?
> I was jus' waitin' for y'all to handcuff me and pat me down
> Touch the detonator in my pocket and we're all goin'
> [BOOM!]
> Are all the pieces comin' together?
> S***, I'm just a crazy sociopath
> that gets off playin' you stupid f***s like a fiddle
> . . . .[91]

At trial, Elonis tried to escape liability for this and his other posts by arguing that they were not threats but instead artistic expression emulating violent rap music.[92] Elonis had even adopted a "rap-style nom de plume" on his Facebook page and had issued "disclaimers that the [posted] lyrics were 'fictitious,' with no intentional 'resemblance to real persons.'"[93] He pointed out that his posts were similar in style and content to those of the rap artist Eminem, an argument that sounds far-fetched unless one actually reads or listens to the lyrics of Eminem's *Kim*.[94]

Elonis also requested that the trial court instruct the jury that "the Government was required to prove that he intended to communicate a 'true threat,'" but the trial court declined.[95] The court instead instructed the jury that they could find a threat if Elonis intentionally made a statement that "a reasonable person would foresee . . . would be interpreted by those to whom the maker communicates the statement as a serious expression of an intention to inflict bodily injury or take the life of an individual."[96] In other words, the instruction directed the jury to apply a doubly objective standard, taking into account how a *reasonable* speaker would predict *reasonable* members of his audience would react to his statement.

---

91. *Id.*
92. *Id.* at 2007.
93. *Id.* at 2005.
94. This segment of lyrics gives a flavor of Eminem's *Kim*:
> Don't you get it bitch, no one can hear you?
> Now shut the fuck up and get what's comin' to you
> You were supposed to love me
> Now bleed! bitch bleed!
> Bleed! bitch bleed! bleed!

EMINEM, *Kim*, *on* THE MARSHALL MATHERS LP (Aftermath, Interscope 2000).
95. *Elonis*, 135 S. Ct. at 2002.
96. *Id.* at 2007.

A jury convicted Elonis on four of the five threats counts against him, and the Court of Appeals upheld his convictions, holding that the federal threats statute required only that a defendant intend "to communicate words that the defendant understands, and that a reasonable person would view as a threat."[97] In other words, the Court of Appeals held that Elonis did not need to subjectively intend to threaten anyone to be convicted of making a threat; it was enough that he negligently made the threat.

The Supreme Court reversed and remanded, but only on statutory construction grounds, not on First Amendment grounds.[98] The federal threats statute under which Elonis was charged provides that "[a]n individual who 'transmits in interstate or foreign commerce any communication containing any threat to kidnap any person or any threat to injure the person of another' is guilty of a felony."[99] The drafters of the statute did not set a mental state for the offense.[100] Nonetheless, the Court relied on precedent, which permits scienter to be presumed when necessary to separate wrongful from innocent conduct,[101] to hold that the "reasonable person" standard used by the lower courts was "inconsistent with 'the conventional requirement for criminal conduct—awareness of some wrongdoing.'"[102] Beyond saying that the statute required a higher standard of scienter than reasonableness, however, the Court left much unresolved.[103] The Court held that if the defendant spoke "for the purpose of issuing a threat, or with knowledge that the communication will be viewed as a threat," such speech would satisfy the statute's requirement.[104] What is unclear, however, is whether a defendant speaking recklessly in making a threat, rather than purposefully or knowingly, would satisfy the federal threats statute's requirements.[105]

As a result, the Court's first social media threats case turned out to be anticlimactic.[106] The Court merely held that under a single federal statute a

97. *Id.*
98. *Id.* at 2008.
99. *Id.* (citing 18 U.S.C. § 875(c) (2012)).
100. *Id.*
101. United States v. X-Citement Video, Inc., 513 U.S. 64, 72 (1994).
102. *Elonis*, 135 S. Ct. at 2011 (emphasis omitted).
103. *Id.* at 2013 (expressly leaving open the question of whether recklessness would satisfy the statute because the parties did not brief it and "[n]o Court of Appeals has . . . addressed that question").
104. *Id.* at 2012.
105. *Id.* at 2013.
106. As Professor Eugene Volokh, a well-respected First Amendment scholar, summarized in the wake of the decision:

> We still don't know, following *Elonis*, whether the "true threats" exception to the First Amendment (1) covers only statements said with the purpose of putting someone in fear, (2) applies also to statements said knowing that the target will be put in fear, (3) applies also to statements said knowing that there's a serious risk that the target will be put in fear, or (4) covers all statements that a reasonable person would view as aimed at putting the target in fear. Indeed, as best I can tell, the Supreme Court did not resolve the federal circuit court disagreement on the First Amendment issue that helped persuade the Court to hear the case.

speaker's negligence would be an insufficient mens rea for conviction. This narrow holding leaves the First Amendment parameters of the true threats doctrine as murky as they were in *Black*.[107] The Court also missed the opportunity to contemplate how social media contexts such as Facebook might affect the interpretation of threatening speech. It is to this issue that we now turn.

II.
UNDERSTANDING THE SOCIAL MEDIA CONTEXTS OF VIOLENT SPEECH

Understanding the varied and sometimes quirky contexts within social media is essential to determining whether remarks made on a social media platform are true threats or protected speech. A number of features of social media contributes to an increase in speech not protected by the First Amendment—including threats, incitement, and harassment—as well as an increase in deeply uncivil but protected speech. Yet some of the same features that make threatening speech more common in social media also magnify the potential for misinterpretation. This Section first explores the set of interrelated features of social media that increase problematic speech and then examines features that make separating the harmful from the hyperbolic more difficult for legal decision-makers.

A. Social Media and Incendiary Speech

In 1968, science fiction author Arthur C. Clarke and film director Stanley Kubrick wrote: "The more wonderful the means of communication, the more trivial, tawdry, or depressing its contents seemed to be."[108] Social media

Eugene Volokh, *The Supreme Court Doesn't Decide When Speech Becomes a Constitutionally Unprotected "True Threat,"* WASH. POST: VOLOKH CONSPIRACY (June 1, 2015), https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/06/01/the-supreme-court-doesnt-decide-when-speech-becomes-a-constitutionally-unprotected-true-threat [http://perma.cc/D2E9-YE3Q].

107. Because the Court's decision in *Elonis* involved statutory construction in one federal threats statute, it did not signal that lower courts had to adopt a subjective test as a matter of constitutional law for all true threats cases. At least two lower court cases decided since *Elonis* maintain that the First Amendment continues to allow an objective standard for judging threats. In *State v. Trey M.*, 186 Wash. 2d 884, 888, *cert. denied*, 138 S. Ct. 313 (2017), the Supreme Court of Washington held that the Supreme Court's *Elonis* decision did not affect its use of an objective test for what constitutes a true threat because "Elonis expressly avoid[ed] any First Amendment analysis" and thus "provide[d] no basis for this court to abandon its established First Amendment precedent." In *United States v. White*, 810 F.3d 212, 220 (4th Cir.), *cert. denied*, 136 S. Ct. 1833 (2016), the Fourth Circuit held that the Supreme Court's *Elonis* decision "does not affect our constitutional rule that a 'true threat' is one that a reasonable recipient familiar with the context would interpret as a serious expression of an intent to do harm"; rather, the *Elonis* decision only affected the statutory requirement of intent. Similarly, in *United States v. Rapert*, 75 M.J. 164, 169 (2016), the US Court of Appeals for the Armed Forces held that it was not inconsistent with *Elonis* to continue to judge the element of whether a threat was communicated from an objective standpoint, at least where a separate element required that the defendant "intended the statements as something other than a joke or idle banter, or intended the statements to serve something other than an innocent or legitimate purpose."

108. ARTHUR C. CLARKE & STANLEY KUBRICK, 2001: A SPACE ODYSSEY 48 (1968).

sometimes seem to bear this out, though perhaps primarily because they create a visible record of how some people really speak.[109] In fact, the informality of much social media speech makes it more akin to chitchat rather than written communication, and this feature leads speakers to post things that they would never contemplate putting in writing in other contexts.[110]

The technology of social media encourages informality. People access Twitter, Facebook, and Instagram through phones or computers, and the technology psychologically distances them from the impact of their words, leading people to say things online that they would never say in person.[111] The disinhibiting effect of the technology is increased further if the speaker also posts anonymously or pseudonymously. Even the speed of communication may foster incendiary speech: speakers respond to provocations before good sense can assert itself.[112] And, of course, no editor stands between speaker and audience to interpose good judgment prior to publication.[113]

The informal, spontaneous, often anonymous, and unmediated discourse common in social media magnifies the potential for incendiary language to cross

109. As Judge Lynn Adelman and Jon Deitrich have written, "It is . . . easy to exaggerate the harm that can come from a new means of communication," and "it is critical that we not succumb to the temptation to weaken our protections of speech based on concerns about terrorists and hatemongers and their use of the internet." *See* Lynn Adelman & Jon Deitrich, *Extremist Speech and the Internet: The Continuing Importance of* Brandenburg, 4 HARV. L. & POL'Y REV. 361, 362–63 (2010).

110. Jacob H. Rowbottom, *To Rant, Vent and Converse: Protecting Low Level Digital Speech*, 71 CAMB. L.J. 355, 359 (2012) (contending that prior to the internet, "[t]he bulk of everyday communications would normally fall below the radar and escape legal sanction").

111. The Internet troll, for instance, is "Cyber Environment Dependent," which means that he requires access to the Internet to engage in passive aggression. Michael Nuccitelli, *Internet Trolls: Cyber Environment Dependent with Sadistic, Psychopathic & Narcissistic Traits*, DARK PSYCHOL. (2014), https://darkpsychology.co/internet-trolls [https://perma.cc/JXC5-CXA4]. "Trolling" refers to the "practice of behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose." Erin E. Buckels, Paul D. Trapnell & Delroy L. Paulhus, *Trolls Just Want to Have Fun*, 67 PERSONALITY & INDIVIDUAL DIFFERENCES 97, 97 (2014). Without the "veil of anonymity" that the Internet provides, trolls wouldn't exist. Nuccitelli, *supra*. "[I]f they could not hide behind their technology, 'they would quickly have their ass kicked for their incessant provocations.'" *Id.*

112. *See* Lyrissa Barnett Lidsky & Thomas F. Cotter, *Authorship, Audiences, and Anonymous Speech*, 82 NOTRE DAME L. REV. 1537, 1575 (2007) (discussing disinhibiting effect of computer mediated communication); Noam Cohen, *Spinning a Web of Lies at Digital Speed*, N.Y. TIMES (Oct. 13, 2008), https://www.nytimes.com/2008/10/13/business/media/13link.html [https://perma.cc/Q2DP-7T56] (providing examples of how quickly lies and hoaxes can spread on the Internet); Terence J. Lau, *Towards Zero Net Presence*, 25 NOTRE DAME J.L., ETHICS & PUB. POL'Y 237, 275 (2011) ("[A]ll Internet users bear some responsibility for knowing how the Internet works, especially the three characteristics of reach, speed, and permanence.").

113. Mary Ann Fitzgerald, *Misinformation on the Internet: Applying Evaluation Skills to Online Information*, 24 EMERGENCY LIBR. 9, 10 (1997) (explaining how misinformation is pervasive on the Internet and that users must filter content for reliability); Donovan A. McFarlane, *Social Communication in a Technology-Driven Society: A Philosophical Exploration of Factor-Impacts and Consequences*, 12 AM. COMM. J. 1 (2010) (explaining how the speed of communication online decreases quality of content).

the line into true threats, incitement, and violence.[114] Social media give hateful speakers ready access to global audiences using a communication platform that is always close at hand.[115] Social media thus give hateful speakers a platform to harass and terrorize targets with seeming impunity and spew vitriol that may spur others to violent actions.[116]

Additionally, the contextual dislocations social media enable—including geographic, cultural, and temporal dislocations—magnify the potential for violent audience reactions to incendiary speech.[117] Speech that is innocuous in one country may be considered blasphemous and provoke violent responses in another;[118] speech that is humorous in one community may be a grave insult in another; and speech that is harmless when posted may provoke violence when viewed.[119]

Social media may also encourage incendiary speech by increasing audience polarization, encouraging and normalizing violent rhetoric, and enabling the

---

114. As one commentator has written, "Social media are analogous to open mikes." Ken Strutin, *Social Media and the Vanishing Points of Ethical and Constitutional Boundaries*, 31 PACE L. REV. 228, 242 (2011).

115. *See* Thomas B. Nachbar, *Paradox and Structure: Relying on Government Regulation to Preserve the Internet's Unregulated Character*, 85 MINN. L. REV. 215, 215 (2000) ("The Internet allows people to communicate quickly, across the globe, and at extremely low cost."). For a definition of social media, see RonNell Anderson Jones & Lyrissa Barnett Lidsky, *Of Reasonable Readers and Unreasonable Speakers: Libel Law in a Networked World*, 23 VA. J. SOC. POL'Y & L. 155, 157 (2016) (citing Andreas M. Kaplan & Michael Haenlein, *Users of the World, Unite! The Challenges and Opportunities of Social Media*, 53 BUS. HORIZONS 59, 61 (2010)).

116. *See* Erwin Chemerinsky, *What the Supreme Court Didn't Decide This Week*, MS. MAG.: BLOG (June 3, 2015), http://www.msmagazine.com/blog/2015/06/03/what-the-supreme-court-didnt-decide-this-week [http://perma.cc/M5F7-DVL3] ("Facebook and other social media have made it much more common for people to make threatening statements that cause others to fear for their safety and even their lives.").

117. *See* Lidsky, *supra* note 23, at 148.

118. For example, the satirical French magazine *Charlie Hebdo*'s publication of a caricature of the Prophet Mohammed triggered hundreds of thousands of protesters to march in predominantly Muslim countries. A terrorist attack on *Charlie Hebdo*'s offices in France followed, in which twelve people were murdered. *See* Cassandra Vinograd et. al., *Charlie Hebdo Shooting: 12 Killed at Muhammad Cartoon Magazine in Paris*, NBC NEWS (Jan. 7, 2015), http://www.nbcnews.com/storyline/paris-magazine-attack/charlie-hebdo-shooting-12-killed-muhammad-cartoons-magazine-paris-n281266 [https://perma.cc/E5ET-GXEH]. In another example, a US-produced film trailer called *Innocence of Muslims* "triggered anti-American sentiment among Muslims in Egypt, Libya and elsewhere in 2012." *See* Dan Levine, *YouTube May Show 'Innocence of Muslims' Film: U.S. Court*, REUTERS (May 18, 2015), https://www.reuters.com/article/us-google-film-ruling/youtube-may-show-innocence-of-muslims-film-u-s-court-idUSKBN0O31R220150518 [https://perma.cc/R78A-PNHP]; *see also* Tsesis, *supra* note 26, at 1152 ("[W]hen threats are posted on the Internet, a billboard, or school blackboard, the object of the message might come across the message later, or not at all, but the forewarning of harm may be no less real.").

119. In 2011, videos of a Quran burning uploaded to YouTube in the United States provoked murderous responses in Afghanistan as a result of their perceived blasphemous character. Enayat Najafizada & Rod Nordland, *Afghans Avenge Florida Koran Burning, Killing 12*, N.Y. TIMES (Apr. 1, 2011), www.nytimes.com/2011/04/02/world/asia/02afghanistan.html [https://perma.cc/7S3Z-NT36]. Such geographical dislocations of online speech are ubiquitous, and yet we rarely hold speakers responsible for the violence that ensues, even if it was foreseeable.

creation of communities united by hate. Much has been written since the 2016 election about the politically polarizing effect of social media on the electorate.[120] This same polarizing effect can lead speakers and their audiences to more and more extreme forms of expression. Hate, unmitigated by voices of reason or social norms of toleration, can reinforce violent impulses within vulnerable audience members.[121] Social media gatherings of like-minded hatemongers can create climates that normalize violent rhetoric and ultimately violent action. Even the potential size of audiences receiving hate-mongering speech matters: the larger the audience, the greater the chance that at least one audience member will respond to a call to violent action. Search technology, such as Google, also aids individuals searching for confirmations of their hateful views and lends encouragement and support to indulge violent impulses.[122]

In effect, the burdens of dealing with the production of incendiary speech in social media are not borne equally: such speech appears to disproportionately target women and people of color, especially those who use social media platforms to speak up against perceived injustice.[123] Indeed, there are numerous

120. *See, e.g.*, Yochai Benkler, Robert Faris, Hal Roberts & Ethan Zuckerman, *Study: Breitbart-led Right-Wing Media Ecosystem Altered Broader Media Agenda*, COLUM. J. REV. (Mar. 3, 2017), http://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php [https://perma.cc/U8XJ-MLWK].

121. Peter Margulies, *The Clear and Present Internet: Terrorism, Cyberspace, and the First Amendment*, 2004 UCLA J.L. & TECH. 4, 33 (2004) ("Lack of mediation is a key ingredient in the production of polarization and concerted violence against innocents to achieve political, cultural, or social aims.").

122. *See* Ryen W. White, *Beliefs and Biases in Web Search*, SIGIR '13 3, 6, 11 (2013) (noting that a study showed that "people seek to confirm their beliefs with their [online] searches and that search engines provide positively-skewed search results, irrespective of the truth," and asserting that this is problematic because "[a]nswers found using search engines can affect action in the world").

123. Professor Danielle Keats Citron has documented this phenomenon extensively in her book, HATE CRIMES IN CYBERSPACE (2014). A 2014 incident called "Gamergate" illustrates this alarming pattern. In the words of the *Washington Post*, GamerGate was a "freewheeling catastrophe/social movement/misdirected lynchmob." Caitlin Dewey, *The Only Guide to Gamergate You Will Ever Need to Read*, WASH. POST (Oct. 14, 2014), https://www.washingtonpost.com/news/the-intersect/wp/2014/10/14/the-only-guide-to-gamergate-you-will-ever-need-to-read [https://perma.cc/WX9B-XZ84]. More specifically, Gamergate was a social media harassment campaign targeting several women in the video game industry, largely via the Twitter hashtag #GamerGate. *Id.* Gamergate exploded after an unusual rift between exes hit social media. *Id.* Independent game designer Zoe Quinn's ex-boyfriend implied in a series of blog posts that she had cheated on him with a writer for a prominent gaming site in exchange for favorable reviews of a game she had developed. *Id.* Angry gamers took to Twitter, Reddit, and 4chan to protest the "ethical breaches in gaming journalism." *Id.* But while Gamergate initially appeared to be about Quinn, the movement quickly expanded to include attacks on other women in the gaming industry. *Id.* Anonymous gamers attacked Anita Sarkeesian after she posted a video series about women and gaming; the attacks caused her to leave her home due to fear of physical harm. *Id.* Jenn Frank, a gaming journalist, and Mattie Brice, a game designer, said they would leave the industry due to harassment that Gamergate inflicted upon them. *Id.* As some in the industry began calling out misogyny in gaming, "a motley alliance of vitriolic naysayers" attacked with threats of death and rape. *Id.* Quinn and several others had to flee their homes "mostly for pointing out that the games industry has a problem with representing women. Keith Stuart, *Zoe Quinn: 'All Gamergate Has Done Is Ruin People's Lives*,*'* GUARDIAN (Dec. 3, 2014), https://www.theguardian.com/technology/2014/dec/03/zoe-quinn-gamergate-interview

examples of social media mobs targeting victims with such extreme speech that they can no longer live a normal life. Even if the likelihood of any threats being carried out is small, targets sometimes permanently alter their lifestyles because of legitimate fear such threats engender.[124] This phenomenon happens in part because within the vast population of social media users, there are "trolls" who receive joy from inducing fear:[125] as researchers have found, "trolling culture embraces a concept virtually synonymous with sadistic pleasure."[126]

This effect of social media discourse on vulnerable groups emphasizes the importance of policing true threats, incitement, defamation, and harassment in social media that make it impossible for victims to live their lives or participate fully in society.[127] However, zeal to eradicate the kinds of threats that deny women and people of color full participation in social media discourse should not tip the balance in criminal law toward punishing mere offensive incivility. Social media daily spawn an appalling number of comments that, in isolation, may appear far more menacing than they might when set in their true contexts. Understanding those contexts is thus key to safeguarding free expression in all of its riotous and sometimes offensive glory without forsaking legitimate protections for vulnerable or victimized online populations.[128]

B. Social Media Misunderstandings

The same characteristics that make social media seem rife with dangerous and destructive speech magnify the potential for a speaker's innocent words to be misunderstood. People flock to social media because it encourages the spontaneity, informality, and intimacy of the spoken word;[129] these features

[https://perma.cc/72YN-WERY]; *see also* Eron Gjoni, *Why Does This Exist?*, ZOE POST (Sep. 12, 2014), https://thezoepost.wordpress.com [https://perma.cc/M735-P24J]; Stephen Totilo, *Untitled*, KOTAKU (Aug. 20, 2014, 6:40 PM), http://kotaku.com/in-recent-days-ive-been-asked-several-times-about-a-pos-1624707346 [https://perma.cc/HMV8-2JQZ].

124. *See generally* CITRON, *supra* note 123.

125. *See* Buckels, *supra* note 111 (defining the practice of "trolling").

126. *Id.* at 98.

127. Critics also suggest that those who control social media platforms, including Google, Facebook, and Twitter, should do more to help uncover those who make such threats and take steps to curtail hateful speech online. *See* Lincoln Caplan, *Should Facebook and Twitter Be Regulated Under the First Amendment?*, WIRED (Oct. 11, 2017, 7:00 AM), https://www.wired.com/story/should-facebook-and-twitter-be-regulated-under-the-first-amendment [http://perma.cc/K7R6-F885].

128. As Raphael Cohen-Almagor explains, in discussing what Karl Popper calls "The Paradox of Toleration," societies should not suppress speech or philosophies that would lead to intolerance so long as they can be kept "in check by public opinion" and "rational argument." RAPHAEL COHEN ALMAGOR, THE BOUNDARIES OF LIBERTY AND TOLERANCE: THE STRUGGLE AGAINST KAHANISM IN ISRAEL 25 (1994).

129. MARCEL DANESI, THE SEMIOTICS OF EMOJI: THE RISE OF VISUAL LANGUAGE IN THE AGE OF THE INTERNET 10 (2015) (observing that "writing has assumed many of the functions of face-to-face (F2F) communication" in the "Internet Age").

explain, in part, why sites like Facebook have drawn billions of users.[130] Yet these same features create immense challenges in separating harmful from harmless speech. The law would never presume to regulate all the threatening or hateful things people say to one another, realizing that speakers often say things in the heat of the moment that are ill-considered, thoughtless, hyperbolic, and often forgotten by both the speaker and the audience within moments.[131] Most social media, however, create a record of such thoughtless speech that can take on an entirely different meaning when read or viewed later. The potential for speech to be read outside of its immediate temporal context, however, is just one of the possible vectors of misunderstanding.[132]

Another cause of misunderstanding is the lack of tonal and other nonverbal cues that signal sarcasm, jests, or hyperbole in oral communications.[133] If an adolescent boy tells a friend, "Don't make me kill you!" in a playful tone, nothing is likely to happen. Within moments, the statement is likely forgotten by all concerned. But if the same adolescent boy posts those words in his Twitter feed, tagging his friend, the potential for misunderstanding the words as a threat is significant, especially if the words reach bystanders who lack a frame of reference for understanding the character or intent of the speaker.

Indeed, the absence of tonal cues has led to the development of abbreviations such as J/K for "just kidding," #hashtags signaling sarcasm or jokes,[134] and even emoticons, emojis, and gifs. Emoticons are a grouping of keyboard characters "used to suggest an attitude or emotion in computerized communications."[135] One of the most popular of these is the smiley face, which

130. *Facebook Company Info*, FACEBOOK, http://newsroom.fb.com/company-info [https://perma.cc/79DX-5X97] (indicating that Facebook has 2.13 billion monthly active users as of December 31, 2017).

131. Social media enable both synchronous and asynchronous communications. "Synchronous forms of digital communications require rapid writing, so that the back-and-forth repartee can be maintained in real time without losing the receiver's attention." DANESI, *supra* note 129, at 11.

132. Historian Milton Mayer coined the term "contextomy" to refer to the practice of selectively taking words from their original context to distort their intended meaning. Matthew S. McGlone, *Deception by Selective Quotation*, *in* THE INTERPLAY OF TRUTH AND DECEPTION: NEW AGENDAS IN COMMUNICATION 54, 55–56 (Matthew S. McGlone & Mark L. Knapp eds., 2010) (attributing the coinage to Mayer).

133. On the importance of nonverbal cues to communication, see generally MARK KNAPP & JUDITH A. HALL, NONVERBAL COMMUNICATION IN HUMAN INTERACTION (7th ed. 2010) (describing how face, voice, and body signals provide context for everyday communications and noting that body positioning and movement, as well as cues observed from the communicating environment itself, can affect the meaning of a communication). *See also* Justin Kruger, Nicholas Epley, Jason Parker & Zhi-Wen Ng, *Egocentrism Over E-mail: Can We Communicate as Well as We Think?*, 89 J. PERSONALITY & SOC. PSYCHOL. 925, 925 (2005) (describing five experiments that show people overestimate their ability to "convey emotion and tone" over email due to the absence "of paralinguistic cues such as gesture, emphasis, and intonation").

134. Dmitry Davidov, Oren Tsur & Ari Rappoport, *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*, COLING 2010: POSTER VOL. 241, 242 (2010).

135. Ghanam v. Does, 845 N.W.2d 128, 133 n.4 (Mich. Ct. App. 2014) ("An 'emoticon' is an icon formed by grouping keyboard characters together into a representation of a facial expression.

is formed by typing a colon together with the close parenthetical symbol [:)] so as to approximate a smiling face. Thus, typing a colon and a close parenthetical symbol together can signal that one is joking. Emojis and gifs are more technologically sophisticated ways of signaling an author's emotions. Emojis are pictographs that often replicate facial expressions,[136] and gifs "are short video clips incorporated into social media posts to express an emotion or make a joke."[137] Both are highly popular[138] with social media users precisely because they lend to the bare text of postings some of the emotional cues of oral communication.[139] Some lawyers have already recognized that legal liability may turn on the interpretation of an emoji,[140] and judges have admitted emojis as evidence of intended meanings in criminal trials.[141]

Yet even emojis can cause misunderstandings. Emojis are not always consistent across platforms.[142] For example, the gun emoji in the title of this Article looks like a space pistol on some platforms and like a revolver on others. Not only do emojis render differently on different platforms, but they can shift even within documents. Indeed, the gun emoji looks like a space pistol in the running header of this Article and a revolver on the title page! Emojis can also

---

Emoticons are used to suggest an attitude or emotion in computerized communications.") (citing RANDOM HOUSE WEBSTER'S COLLEGE DICTIONARY (2003)).

136. Emojis are pictographs used to signal emotions in online conversations. *See* DANESI, *supra* note 129, at 10 (contending that the "intent" of emojis is "to add what can be called 'visual tone' to a message" and also noting that they are "used largely in informal communications, to add visual annotations to the conceptual content of a message").

137. Eric Goldman, *Surveying the Law of Emojis*, SANTA CLARA U. LEGAL STUD. PAPER SERIES NO. 8-17, at 12 (May 1, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2961060 [https://perma.cc/3TRJ-7L66].

138. DANESI, *supra* note 129, at 33 ("According to a 2015 study by the British app developer, SwiftKey, which collected and examined over a billion bits of data from Android and iOS devices in sixteen languages, 45 percent of all messages contained happy-face emoji, such as smileys, followed by sad faces, hearts, hand gestures, and romantic emoji."). Another study found that more than 90 percent of social media users employ emojis. *See* Emogi Research Team, *2015 Emoji Report*, EMOGI (Sept. 2015), http://cdn.emogi.com/docs/reports/2015_emoji_report.pdf [https://perma.cc/Z6DF-CKPS].

139. The Oxford English Dictionary named emoji its word of the year in 2015. Heather Kelly, *Emoji Named Word of the Year*, CNN TECH (Nov. 16, 2015), http://money.cnn.com/2015/11/16/technology/emoji-oxford-word-of-the-year [https://perma.cc/E376-BS2A].

140. For an example of an emoji featuring in a defamation case, see *Ghanam*, 845 N.W.2d at 145. There, the court determined that the allegedly defamatory statement "on its face cannot be taken seriously as asserting a fact. The use of the ':P' emoticon makes it patently clear that the commenter was making a joke. As noted earlier, a ':P' emoticon is used to represent a face with its tongue sticking out to denote a joke or sarcasm. Thus, a reasonable reader could not view the statement as defamatory." *Id.*

141. *See* DANESI, *supra* note 129, at 17 (discussing the admission into evidence in a criminal trial of a defendant's use of a smiley-face emoji as negating criminal intent).

142. Goldman, *supra* note 137, at 9 ("Because platforms differ in their implementations of emojis, a Unicode-defined emoji will often appear differently across platforms."). Goldman's excellent article discusses in depth the technology of emojis and the problems they create for legal decision-makers in a variety of contexts. *See generally id.*

lead to misunderstanding because they lack a standard meaning.[143] In a job posting for an "emoji translator," a London firm notes that this emoji, 😂 , is subject to misunderstanding because in the United States it means "laughing so hard I'm crying," but in the Middle East it can be mistaken for grief.[144] As Professor Eric Goldman points out, these interpretive issues are made even more complicated because there are no credible dictionaries of emojis.[145] Gifs, which often consist of small clips from popular television shows or movies, may be even harder to decode than emojis: not only do they usually contain more "content," but correct interpretation often requires underlying knowledge of the show or movie from which they are borrowed.[146] These examples illustrate that speech between "insiders"—those who understand a shared decoding system—can easily go awry when decoded by "outsiders" to the conversation, a decoding problem that social media make more common.

Another feature that leads to misinterpretations in social media is that different social media "sites" have different discourse conventions. Much social media discourse is governed by the norms of informal spoken conversation,[147] an aspect that is sometimes signaled by the presence of frequent typographical errors, misspellings, poor grammar, and profanity. Thus, it is easy to misconstrue speech in social media if one applies interpretive norms applicable to formal written communications. But the interpretive problems run even deeper, because discourse conventions may vary within a single social media platform, and subgroups within those platforms may use the platform differently.[148] Discourse on Facebook tends to be relatively civilized compared to discourse on Reddit,[149] but even within Facebook the way fifteen-year-old teens speak to each other is unlikely to be the same way that forty-eight-year-old lawyers speak to each other. Similarly, some discussion forums within a single social medium, such as Reddit, are much more civilized than others.[150] Obviously, the medium and forum within which a message appears, as well as the demographics and discourse conventions of users within that medium and forum, affect how to

143. *See* Garreth W. Tigwell & David R. Flatla, *"Oh That's What You Meant!": Reducing Emoji Misunderstanding*, MOBILEHCI '16 PROC. OF THE 18TH INT'L CONF. ON HUM.-COMPUT. INTERACTION WITH MOBILE DEVICES & SERVS. ADJUNCT 859 (2016) (demonstrating via a study based on questionnaires that significant variability exists in interpreting the meaning of emojis); *see also* Goldman, *supra* note 137, at 17.

144. *See* Petroff, *supra* note 11; *see also* DANESI, *supra* note 129, at 31 (noting that the "thumbs up" emoji represents a gesture that "is hideously offensive in parts of the Middle East, West Africa, Russia, and South America").

145. *See* Goldman, *supra* note 137, at 17.

146. *See id.* at 12–13.

147. *See generally* Jones & Lidsky, *supra* note 115, at 166–67.

148. *See* Kim Holmberg & Mike Thelwall, *Disciplinary Differences in Twitter Scholarly Communication*, 101 SCIENTOMETRICS 1027 (2014) (discussing the varied ways that scholars within different disciplines use Twitter).

149. CARRIE JAMES, DISCONNECTED: YOUTH, NEW MEDIA, AND THE ETHICS GAP 116 (2014).

150. *Id.* at 96–97.

interpret the message. And yet these contextual features are likely to be opaque to those who have not spent time in the mediums or forums.

At a still deeper layer of online context, every individual interaction may develop its own conventions on the spot. We speak differently to strangers than to friends, differently to business associates than to family members, differently to lovers than to mail couriers.[151] And social media use has become such an engrained practice for so many speakers that they forget they are being observed.[152] The disinhibiting effect of the technology encourages us to speak to our Facebook friends as if they are our therapists, with potentially unfortunate consequences for us and those around us.

Even the architectural features of different social media platforms can become an aspect of context that affects correct interpretation. Twitter limits posts to 280 characters,[153] which limits a speaker's ability to signal sarcasm, hyperbole, or jests. A single tweet, considered outside the string of which it is a part, may appear threatening in isolation. Perhaps a tweet that includes a hyperlink to another page (e.g., a blog post), considered outside the context of the content it refers to, could also seem to pose a threat. Similarly, Snapchat's architecture limits the length of time any speaker's message or photo is displayed, which could make an alleged threat seem more or less ominous, depending on other aspects of the surrounding context.

For instance, in January 2017, one high school student's arrest highlighted the problem in trying to assess threats on Snapchat due, in large part, to the platform's distinct features. Indianapolis police arrested a student when a seventeen-year-old classmate reported receiving a threat on Snapchat directed at

151. *See, e.g.*, Daantje Derks, Arjan E.R. Bos, & Jasper von Grumbkow, *Emoticons and Social Interaction on the Internet: The Importance of Social Context*, 23 COMPUTS. IN HUM. BEHAV. 842, 846 (2007) (explaining that because social norms make it more appropriate to show one's emotions towards friends than towards colleagues, people online tend to use more emoticons in "socio-emotional contexts" than in "task-oriented" contexts).

152. Psychologists hypothesize that this is due to the apparent "distance" between the speaker and the audience. Noam Shpancer, *Why You Might Share More Intimately Online*, PSYCHOL. TODAY (June 24, 2014), https://www.psychologytoday.com/blog/insight-therapy/201406/why-you-might-share-more-intimately-online [https://perma.cc/E3G6-4B3L]. Computer-mediated communications tend to lead to over-sharing of personal, intimate details due to the sense of security the Internet seems to offer. *Id.* Paradoxically, this increase in self-disclosure online may also be due to the absence of nonverbal, visual cues online, because that absence also creates heightened uncertainty, which people seek to reduce by disclosing more. *Id.*

153. Twitter initially limited tweet length to 140 characters, so that tweets could be sent in text messages with 160-character limits, but subsequently expanded the limit to 280 characters for almost all users. *See* Selena Larson, *Welcome to a World with 280-Character Tweets*, CNN TECH (Nov. 7, 2017, 4:00 PM), http://money.cnn.com/2017/11/07/technology/twitter-280-character-limit/index.html [https://perma.cc/GKC7-98WV] ("The new character length won't apply to Japanese, Korean or Chinese-language tweets. Those languages can convey twice as much information in less space, so tweets will remain at 140 characters, Twitter said.").

the student's school.[154] After authorities identified the student who sent the message, the school released the following statement:

> Because of the nature of Snapchat, the content of the threat remains unclear to school officials. After [police] conducted an investigation, they were able to track down the individual who was responsible and the student was arrested for making the threat. . . . [I]n an abundance of caution [we] have requested an increased police presence at school to reassure students, staff and parents that we take all threats seriously.[155]

Neither the school nor law enforcement could properly interpret the nature of the student's original "threat" because Snapchat messages disappear moments after they have been opened—one of the platform's main attractions to its users—forcing them to rely on the memory (and possibly skewed perception) of the student who reported the threat. While there are ways to retrieve the deleted messages,[156] the process requires time—something a school official responding to a threat cannot afford. When dealing with uncertainty about the threat itself on the one hand, and the safety of children on the other, the threat will likely lead to overreaction, fear, and a risk of criminal consequences for the speaker, regardless of whether the threat was made in jest or sarcastically.[157]

Cases like this also illustrate the generation gap problem inherent to social media. While platforms like Snapchat, Tumblr, and Instagram have the youngest audiences, with about 40 percent or more of their users coming from the teenage or young adult demographic, older audiences seem to flock towards platforms like Facebook and LinkedIn.[158] The generation gap within certain social media platforms—especially when combined with the distinct communication

---

154. FOX59 Web, *Emmerich Manual Student Accused of Making Online Threats Directed at School Arrested*, FOX 59 (Jan. 13, 2017), http://fox59.com/2017/01/13/arrest-made-after-online-threats-directed-at-emmerich-manual-high-school [https://perma.cc/9XFU-SRGB].

155. *Id.*

156. *Id.*

157. Another 17-year-old, Jashon Jevon Taylor, was charged with the felony of making a terrorist threat. *See* Max Londberg, *Belton High Teen Charged with Felony After Threat on Snapchat*, KANSAS CITY STAR (Feb. 8, 2017), http://www.kansascity.com/news/local/crime/article131622874.html [https://perma.cc/M5TS-5K39]. Angry that his team did not win the Super Bowl, he took to Snapchat to express his hatred towards the New England Patriots and their fans, stating, "Let it be know im shooting Belton tomorrow or tuesday . . . Imma find every (person) bragging and blast them." *Id.* It is unclear whether Taylor's threat was intended to be taken seriously, considering his age and the larger context of sports rivalries, which often inspire baseless threats between opposing teams (e.g., "We're going to kill you," "we're going to crush you," etc.). But if, as suspected, Taylor's comments were mere hyperbole, considering the surrounding context would be important before finding him guilty on a felony charge.

158. *Social Summary: GlobalWebIndex's Quarterly Report on the Latest Trends in Social Networking*, GLOBALWEBINDEX 1, 10 (2016), http://insight.globalwebindex.net/hubfs/Reports/GWI-Social-Q4-2016-Summary-Report.pdf?utm_campaign=Flagship%20Reports&utm_source=hs_automation&utm_medium=email&utm_content=38360477&_hsenc=p2ANqtz--MIbnoOLhnfH1DiiAO0F3zRHD1DTm2pLp6SpLPoBx3n_IJgaAgyBU57J3mSRcCdjYhNsAg5xtz2dOZN1uiuvrG6buySA&_hsmi=38360477 [https://perma.cc/53ZM-5FV2].

conventions that develop within each of them—may lead courts and lawmakers unfamiliar with those conventions to criminalize normal or common adolescent behavior, which has increasingly included the use of hyperbole in almost any given online situation.[159]

For instance, if anyone were to read the comments on a photo of a cute animal posted to Instagram, the observer could legitimately think that the commenters, comprised of mostly girls and young women, were facing certain extinction because many seem to be "literally dying."[160] But of course, "literally" here really means "figuratively," and "dying" does not really mean they have met their demise. As one author's twenty-year-old research assistant put it, "It's almost like 'dying' has become a filler for anytime anyone says anything remotely entertaining. . . . Like, if what you're saying won't legitimately put me to sleep, I respond with, 'OMG dying.'"[161] This type of hyperbolic death can also take the form of:

> "dying" (death in process), "not breathing" (first sign of possible death), "all the way dead," "actually dead" and "literally dead" (just so you know), as well as "literally bye" (for when you're about to die), "ded" (when you are dying so fast that typing an "a" would delay the entire process) and "RIP me" (after you've had a moment to process it). There's also kms, or "killing myself," which, as 15-year-old Ruby Karp, a high school student in Manhattan, explained it, can be used to say something like "ugh so much homework kms!"[162]

This is a generational convention that has developed among teens and millennials through the use of computer-mediated communication. It is no wonder then that when an older "outsider" unfamiliar with the lingo becomes privy to the conversation, she may tend to overreact to a perceived threat of suicide. Imagine a dad posts a baby picture on Facebook and his daughter comments on the photo with a single word: "dead." Someone familiar with this type of hyperbole, now so predictable from the daughter's age group, would understand that the daughter means, "Dead at that pic cause it's rly cute!!!"[163] But the comment could just as easily force the dad into a state of panic, concluding that his daughter is having some kind of psychological breakdown or needs to be put on suicide watch.

While social media encourage hyperbole through the disinhibiting effect of anonymity, this type of "death" hyperbole and over-exaggeration that have become natural functions of conversation between young people is due, paradoxically, to the performative element also integral to social media use.[164]

---

159. *See* Jessica Bennett, *OMG! The Hyperbole of Internet-Speak*, N.Y. TIMES (Nov. 28, 2015), https://www.nytimes.com/2015/11/29/fashion/death-by-internet-hyperbole-literally-dying-over-this-column.html [https://perma.cc/99JL-8LSN].

160. *See id.*

161. *Id.*

162. *Id.*

163. *Id.*

164. *Id.*

People post things in social media knowing that there is an audience.[165] It is perhaps this function of social media that has led to the inescapable edited-selfie culture.[166] To boost their personal brand by getting those "likes," comments, or shares, they have to perform; they have to be interesting.[167] Hyperbole and exaggeration is one way to garner that attention, but as already illustrated, not everyone on the other side of the screen may have the same understanding of the true meaning or intent behind the words.

Given the many misunderstandings that can arise in social media contexts, the challenge is to create legal rules and procedures that allow for suppression of threats while leaving ample breathing room for everyday social media conversation in all its mundane, often profane, and hyperbolic glory.

III.
TRASH TALK OR TRUE THREAT? WHY IT MATTERS

As the preceding discussion demonstrates, it is easy for "outsiders," including legal decision-makers, to misconstrue speech in social media. Such misconstructions may include interpreting violent hyperbole as true threats. Some would contend, however, that such misconstructions are of little concern to criminal law or the First Amendment.[168] Consider again the case of Justin Carter. Carter made a hyperbolic, uncivil, and deeply offensive comment that made light of the murder of schoolchildren. Even if he and his immediate audience viewed his comments as hyperbole rather than a true threat, one might argue that speakers like Carter should assume the risk of being misunderstood.[169] Threats, after all, have been branded "low-value" speech by the Supreme Court

165. Some even become addicted to validation from that unseen audience. *See* Maureen O'Connor, *Addicted to Likes: How Social Media Feeds Our Neediness*, NY MAG. (Feb. 20, 2014), http://nymag.com/thecut/2014/02/addicted-to-likes-social-media-makes-us-needier.html [https://perma.cc/HQ33-VGFY].

166. < *Narcissistic, Maybe. But Is There More to the Art of the Selfie?*, ALL TECH CONSIDERED (NPR broadcast July 27, 2015) (transcript available at http://www.npr.org/templates/transcript/transcript.php?storyId=425681152 [https://perma.cc/WGU5-P9P2]).

167. *See* O'Connor, *supra* note 165.

168. Professor Eugene Volokh contends that speech is most problematic, and least protected by the First Amendment, when it is "'unwanted one-to-one speech'—speech said to a particular person in a context where the recipient appears not to want to hear it, whether because the recipient has expressly demanded that the speech stop or because the speaker intends to annoy and offend the recipient." Eugene Volokh, *One-to-One Speech vs. One-to-Many Speech, Criminal Harassment Laws, and "Cyberstalking*,*"* 107 NW. U.L. REV. 731, 742 (2013).

169. Forcing Carter and other speakers to assume the risk of being misunderstood would prioritize order and civility at the expense of free expression. It is true that the doctrinal boundaries set by criminal law help maintain the minimal level of civility and security necessary for participatory public discourse—in social media and elsewhere. Social media companies also play a role in setting boundaries so discourse can thrive, though some have predicted that social media will fracture into "safe spaces" patrolled by artificial intelligence. Lee Rainie et al., *The Future of Free Speech, Trolls, Anonymity and Fake News Online*, PEW RES. CTR. (Mar. 29, 2017), http://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online [https://perma.cc/DS65-3HP8].

because, like fighting words, libel, and obscenity,[170] they are of "such slight social value as a step to truth that any benefit that may be derived [is] clearly outweighed by the social interest in order and morality."[171]

But just because threats are in the "low-value" category of speech does not mean that line drawing may be any less precise with regard to threats than it can with regard to higher-value categories.[172] The history of the Supreme Court's use of "low-value" categories over the last half century or so has been an attempt to narrowly define them to leave ample "breathing space"[173] so that protected speech is not suppressed. For example, the Court has recognized that the labels "libel" and "obscenity" are not self-defining, and the Court's decisions have carved out some "breathing space" around protected speech, even at the risk of not punishing some speech that is libelous or obscene.[174] Moreover, the Court has barred states from punishing even low-value speech based on opposition to its ideological content, as opposed to its dangerous or destructive content or qualities.[175] In other words, the Court has recognized that clearing up doctrinal uncertainty surrounding low-value categories is necessary for free speech to thrive.

Furthermore, it is important to remember that all speech that is not squarely within one of the Court's categories of low- or intermediate-value speech presumptively lies at the core of the First Amendment. So-called core speech

---

170. *See* Roth v. United States, 354 U.S. 476, 483 (1957) (holding that obscenity is "outside the protection intended for speech and press"); Beauharnais v. Illinois, 343 U.S. 250, 266 (1952) (holding that libel is not "within the area of constitutionally protected speech"); Chaplinsky v. New Hampshire, 315 U.S. 568, 572 (1942) (holding that fighting words are unprotected).

171. *Chaplinsky*, 315 U.S. at 571–72 (noting that the "prevention and punishment" of these categories "have never been thought to raise any Constitutional problem"); *see also* United States v. Stevens, 559 U.S. 460, 469 (2010) (noting that the existence of a "long-settled tradition of subjecting that speech to regulation" is required to establish a low-value category). For an excellent article questioning the historical basis for the "low-value" categories, see Genevieve Lakier, *The Invention of Low-Value Speech*, 128 HARV. L. REV. 2166 (2015). *See also* Frederick Schauer, *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 HARV. L. REV. 1765, 1769 (2004) (explaining that in constitutional doctrine some things that are speech are simply defined as not speech—speech incident to criminal conduct is conduct and not speech and therefore receives no protection).

172. *See* Rowbottom, *supra* note 110, at 369–70 (arguing, for example, that even casual conversation is a way in which speakers "decide how to present [themselves] to society" and "mak[e] social connections, and people's reactions to such comments will provide a route to discovering social norms"). Rowbottom contends that this argument provides support, though limited, for protecting such speech. He ultimately argues for "proportionate" restrictions on "day-to-day conversations" on the internet. *Id.*

173. N.Y. Times v. Sullivan, 376 U.S. 254, 272 (1964) (citing N.A.A.C.P. v. Button, 371 U.S. 415, 433 (1963)).

174. As Professor Michael Birnhack has observed with regard to obscenity, "Obviously, the difficulty lies in drawing the line between the two kinds of content—a problem with which American courts struggle. This difficulty in itself has a price—the unclear boundaries of the 'illegal' might deter not only illegal speech, but also legitimate content." Michael D. Birnhack & Jacob H. Rowbottom, *Shielding Children: The European Way*, 79 CHI.-KENT L. REV 175, 183 (2004).

175. R.A.V. v. City of St. Paul, 505 U.S. 377, 382–83 (1992).

includes discussion of political, literary, artistic, historical, cultural, and social concerns,[176] but it also includes speech that, put simply, the government has no justifiable reason to regulate. When a law professor complains about grading student papers on Facebook, she is likely not contributing to democratic self-governance, but neither is the government free to ban such communication. Similarly, when a law student posts that he hates his Torts professor and wishes she would drop dead, the government cannot jail the speaker because the speech appears to make little contribution to the student's self-realization.

As Professor Clay Calvert has written, "First Amendment protection for speech is not based or dependent on some abstract, qualitative judicial judgment about how much literary or societal value that speech holds."[177] Rather, it is based on the a priori judgment that the government has no business regulating speech that does not cause harm.[178] The First Amendment protects deeply offensive, racist, or misogynistic comments not because they provide substantial value to public discourse, nor even because they do no harm. Rather, they are protected because they do no harm cognizable by the First Amendment. The harms such speech cause typically can be remedied by counterspeech, and the perils of allowing the government to regulate this type of uncivil discourse are too dangerous to unleash.[179] This, then, is why precision in line drawing is

176. *See* Harry Kalven, Jr., *The New York Times Case: A Note on "The Central Meaning of the First Amendment,"* 1964 SUP. CT. REV. 191, 208 (1964) ("The Amendment has a 'central meaning'—a core of protection of speech without which democracy cannot function, without which, in Madison's phrase, 'the censorial power' would be in the Government over the people and not 'in the people over the Government.'").

177. Clay Calvert & Robert D. Richards, *The 2003 Legislative Assault on Violent Video Games: Judicial Realities and Regulatory Rhetoric*, 11 VILL. SPORTS & ENT. L.J. 203, 215 (2004); *see also* Calvert & Bunker, *supra* note 18, at 960–61 (critiquing the Supreme Court's failure to clarify true threats doctrine); Clay Calvert, Emma Morehart & Sarah Papadelias, *Rap Music and the True Threats Quagmire: When Does One Man's Lyric Become Another's Crime?*, 38 COLUM. J.L. & ARTS 1, 1 (2014) (analyzing the "complex and unsettled state of the true threats doctrine through the lens of the equally complicated, controversial and multi-faceted musical genre of rap").

178. S. Elizabeth Wilborn Malloy & Ronald J. Krotoszynski, Jr., *Recalibrating the Cost of Harm Advocacy: Getting Beyond* Brandenburg, 41 WM. & MARY L. REV. 1159, 1186 (2000) (describing the unprotected speech category as "unprotected precisely because the threatened social harms associated with the speech outweigh any potentially offsetting social value associated with the particular type of speech").

179. C. Edwin Baker, *Scope of the First Amendment Freedom of Speech*, 25 UCLA L. REV. 964, 966 (1978) (arguing that under the "liberty model," "the free speech clause protects not a marketplace but rather an arena of individual liberty from certain types of governmental restrictions").

required even in low-value categories,[180] and even for speech like Justin Carter's, which does not seem to make a significant contribution to public discourse.[181]

Such speech is nonetheless protected where the government lacks a permissible justification for restricting speech,[182] and the only permissible justifications are not just harm, but harm that cannot be remedied effectively by measures short of government restriction, including counterspeech.[183] This point

180. Some have tried to argue that speech containing violent imagery falls into a different category. Yet even violent speech is protected under the First Amendment, and the Supreme Court has explicitly rejected the argument that violent expression may be restricted lest it lead to violent action. In *Brown v. Entertainment Merchants Ass'n*, 564 U.S. 786 (2011), the state of California sought to restrict the sale and rental of violent video games to children, citing concerns that exposure to interactive violence is linked to antisocial and violent behavior. *Id.* at 789. The state asserted that the immersive interactivity of violent video games made them so dangerous to minors that they were analogous to obscenity and could be placed into an unprotected category of speech. *Id.* at 793. The Supreme Court rejected the argument. *Id.* at 792. Although the Court expressed some skepticism about the value of video games to public discourse, the Court nonetheless accorded the games full constitutional protection, stating: "[W]e have long recognized that it is difficult to distinguish politics from entertainment, and dangerous to try." *Id.* at 790. The state's attempts to regulate the games failed in significant part because the Court was unconvinced that watching or reading violent entertainment produces violent action, and the state failed to prove any causal link between them. *Id.* at 798–99. The Court found the legislation particularly unnecessary in light of the evidence on the record that the video game industry already performed a great deal of self-policing by voluntarily refusing to sell to minors video games (voluntarily) labeled as appropriate only for mature audiences. *Id.* at 803. Furthermore, the Court was unwilling to open a new category of protected speech merely because a new technology made the "speech" seem more dangerous. *Id.* at 790. Absent "persuasive evidence" that society had long condemned this type of speech, "a legislature may not revise the 'judgment [of] the American people,' embodied in the First Amendment, 'that the benefits of its restrictions on the Government outweigh the costs.'" *Id.* at 792 (quoting United States v. Stevens, 559 U.S. 460, 470 (2010)). Clearly, entertainment that might be construed as containing threats is common. For example, the following songs might be construed as threatening: Cop Killer's *Body Count*, The Beatles's *Run for Your Life*, Foster the People's *Pumped Up Kicks*, Pearl Jam's *Jeremy*, NWA's *Fuck the Police*, Carrie Underwood's *Before He Cheats*, Nirvana's *Where Did You Sleep Last Night?*. Clearly some violent rhetoric in social media is distinguishable from violent video games or even violent lyrics and imagery in popular music. Justin Carter's argument with his interlocutor on Facebook presumably was not intended to be a form of mass entertainment for a large audience, and even though Anthony Elonis claimed his speech was just such an entertainment, that claim is highly dubious given the tone, setting, and factual background.

181. See Andrew Koppelman, *Madisonian Pornography or, the Importance of Jeffrey Sherman*, 84 CHI.-KENT L. REV. 597, 608 (2009) (observing that "free speech theory protects even worthless and harmful speech") (citing George Kateb, *The Freedom of Worthless and Harmful Speech*, *in* LIBERALISM WITHOUT ILLUSIONS: ESSAYS ON LIBERAL THEORY AND THE POLITICAL VISION OF JUDITH N. SHKLAR 220–40 (Bernard Yack ed., 1996)).

182. *See* Schauer, *supra* note 73, at 207.

183. The Court in *United States v. Alvarez*, 567 U.S. 709, 729 (2012), reaffirmed that the First Amendment protects even relatively valueless "speech we detest" when the harms the speech causes can be "cured" by counterspeech or other measures. *Alvarez* struck down a federal statute making it a crime for a person to falsely claim that she received a military decoration or medal authorized by Congress. *Id.* at 729–30. Though the *Alvarez* Court was divided, the decision affirmed that the government lacks the power to censor lies—even lies that offend patriotic values—absent a showing of significant harm. *Id.* at 719. The holding in *Alvarez* represents a sign that it matters what order we ask the questions: Does this speech have value? Does it cause harm? And would government regulation of it cause more harm than the speech itself? The Court asked the "harm questions" before the value question: although the speech was relatively valueless, the government could nonetheless not suppress

requires some elaboration. The true threats doctrine is based on the prevention of constitutionally cognizable harms—the harm of putting people in fear, the disruption that fear engenders, and the disruption and harm of violence itself.[184] Arguably, these first two harms manifest when a threat is taken seriously by a target, or by someone charged with protecting the target; one might extrapolate, therefore, that any time social media speech causes *fear* it has "inflict[ed] injury by [its] very utterance" and can be suppressed.[185] Yet the emotion of fear is so variable and so difficult to predict that the fact speech puts *someone* in fear, standing alone, cannot and should not be the basis for criminal liability.[186] If protecting citizens from fear were the only value we sought to protect, we would simply view threats solely from the subjective perspective of the target of the threat or the subjective perspective of individual audience members.[187] And if this were the case, even crime and disaster news, violent programming, or fictional horror movies could be censored through criminal sanctions because they both reflect and provoke worries and fear of a violence-filled world.[188]

But both common law and constitutional law have recognized that tort liability (much less criminal liability) cannot be based solely on causing someone the subjective emotion of fear. In tort law, speech cannot be the basis for intentional infliction of emotional distress liability unless it inflicts "severe" (that is, extreme and prolonged) emotional distress.[189] In addition, the speaker must intentionally or recklessly inflict distress.[190] Moreover, the conduct must be such as to be "utterly intolerable in a civilized community."[191]

Even then, the Supreme Court has imposed further limits on the imposition of tort liability for emotional harm caused by speech. A plaintiff may not recover for intentional infliction of emotional distress caused by speech that involves a

---

it because the harm it caused was slight and could be remedied by simpler, less speech-restrictive means than criminalization. *Id.* at 726, 729.

184. Virginia v. Black, 538 U.S. 343, 360 (2003).

185. *See* Schauer, *supra* note 73, at 214.

186. *See* Avlana K. Eisenberg, *Criminal Infliction of Emotional Distress*, 113 MICH. L. REV. 607, 610 (2015) (explaining why criminal law traditionally has been reluctant to criminalize emotional distress); *see also* Christina Wells, *Regulating Offensiveness:* Snyder v. Phelps*, Emotion, and the First Amendment*, 1 CALIF. L. REV. CIR. 71, 72–73 (2010) (discussing the variability of human emotions in response to speech and the dangers of using emotional responses as a basis for censorship).

187. *See* Eisenberg, *supra* note 186, at 618–20.

188. George Gerbner & Larry Gross, 26 J. COMM. 172, 191–94 (1976).

189. Restatement (Second) of Torts § 46 (A.L.I. 1965) (setting forth elements of tort of intentional infliction of emotional distress).

190. *Id.*; *see also* State Rubbish Collectors Ass'n v. Siliznoff, 240 P.2d 282, 284 (Cal. 1952). *Siliznoff* involved a garbage collector who was threatened with physical harm and property damage unless he paid a bribe to rival garbage collectors. *Id.* Even though the threats involved future harms, and thus were not assaults, the court concluded that "a cause of action is established when it is shown that one, in the absence of any privilege, intentionally subjects another to the mental suffering incident to serious threats to his physical well-being, whether or not the threats are made under such circumstances as to constitute a technical assault." *Id*. at 284–85.

191. Restatement (Second) of Torts § 46, cmt. d (A.L.I. 1965).

matter of public concern.[192] Nor may a plaintiff who is a public figure recover for emotional harm inflicted by a parody or satire, absent proof that the parody implied false facts made with knowledge or reckless disregard of their falsity.[193] These elaborate safeguards suggest that both tort and constitutional law recognize that protecting society from speech that causes fear or distress must be balanced against other values, including the danger of suppressing valuable or at least "innocent" speech.

The balance struck by criminal law between free speech and emotional security should be no less speech-protective than the balance struck by tort law. If anything, the First Amendment should protect speakers even more from criminal than tort liability.[194] At a minimum, the First Amendment's protection should shield speakers from criminal liability imposed without proof that the speaker's purpose was to make her target reasonably fear violence or that the speaker knew with substantial certainty that her words would make the target reasonably fear violence.[195] The requirement of specific intent provides necessary insurance that threats that were not taken seriously by reasonable people within the speaker's immediate audience will not lead to liability. In other words, the requirement of specific intent provides some insurance against a speaker being punished for speech taken out of context. Yet in order to truly protect free speech, context must be considered in an even more robust fashion.

Again, the intersection of the First Amendment and tort law is instructive. In the tort of defamation, the Supreme Court has recognized that context is a critical determinant of whether a statement is defamatory and has directed lower courts to adopt the perspective of the reasonable reader[196] in determining whether

192. *See* Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 50 (1988) (holding public figures may not obtain damages for emotional distress caused by speech unless speech was false and made with actual malice); Snyder v. Phelps, 592 U.S. 443, 458 (2011) (holding, in part, that First Amendment protection of speech does not disappear with a jury verdict finding the conduct was "outrageous").

193. *Hustler*, 485 U.S. at 56.

194. Curiously the Supreme Court has tended to interpret the First Amendment as imposing the same boundaries on the reach of tort and criminal law. See, for example, *Florida Star v. B.J.F.*, 491 U.S. 524 (1989), a case brought by plaintiff as a negligence per se action. There, the Supreme Court imposed the same First Amendment limits on tort law that it had used previously in a case involving a criminal statute. *Id.* at 536 (citing Smith v. Daily Mail Publ'g Co., 443 U.S. 97, 103 (1979)). Professor David A. Anderson criticized this development in his article, *First Amendment Limitations on Tort Law*, 69 BROOKLYN L. REV. 755 (2004). The argument that the First Amendment imposes more limits on criminalizing speech than it does on imposing tort liability is deserving of further development in the scholarly literature.

195. *See generally* Schauer, *supra* note 73, at 217.

196. The reasonable reader is a legal construct—a hypothesized reader who is a sophisticated decoder of the contextual clues provided to reach the meaning that social norms suggest she should reach. *See* David McCraw, *How Do Readers Read? Social Science and the Law of Libel*, 41 CATH. U.L. REV. 81, 104 (1991). The reasonable reader of social media texts should be one who decodes them similarly to sophisticated actual readers—those aware of the discourses and conventions within the medium and the technological architectures that may alter meaning. *Id.* at 99. Determining meaning according to this hypothesized reasonable reader protects important public policy interests, including safeguarding the vitality of discourse both in traditional and new media of expression. *See* Lyrissa Barnett Lidsky, *Nobody's Fools: The Rational Audience as First Amendment Ideal*, 2010 U. ILL. L.

speech is capable of a defamatory meaning. Further, the Court has held that the First Amendment bans liability for defamation based on satire, parody, hyperbole, and other types of figurative speech that are either not provably false or cannot reasonably be interpreted as stating actual facts.[197] Determining whether statements fall into these categories requires close consideration of the words used, but it also requires sophisticated consideration of all of the surrounding context. Indeed, some lower courts both before and after *Milkovich* have specified the types of contextual factors that should be considered in determining whether an alleged defamatory statement implies a factual assertion.[198] The Ninth Circuit, for instance, specified that legal decision-makers must look at the "totality of circumstances"[199] and specified three factors in particular that must be examined in interpreting whether a statement is defamation, namely:

> (1) the statement in its broad context, which includes the general tenor of the entire work, the subject of the statements, the

REV. 799, 842 (2010) (noting that the Supreme Court has "clearly endorsed the principle that speakers should not be held liable for 'mis-readings' of their speech by idiosyncratic or unsophisticated audience members" because imposing such liability would leave insufficient breathing space for free expression). In *FEC v. Wisconsin Right to Life, Inc.*, 551 U.S. 449, 469–70 (2007), Chief Justice Roberts, joined by Justice Alito, explained that the First Amendment requires the line between protected and unprotected political speech to be drawn based on a reasonable interpretation of what the effect on the audience was likely to be rather than the actual effects. *Id.* at 469–70. Otherwise, the search for empirical evidence of "actual effects" would be likely to "chill a substantial amount of political speech." *Id.* at 468–69.

197. Milkovich v. Lorain Journal Co., 497 U.S. 1, 17, 25 (1990).

198. *Milkovich* did not specify all of the factors that courts might consider in determining whether a statement is defamatory. As a result, it was unclear whether the Supreme Court in *Milkovich* was rejecting the multifactored tests some courts had used previously to address the issue in cases such as *McCabe v. Rattiner*, 814 F.2d 839, 842 (1st Cir. 1987) (applying totality of the circumstances analysis) and *Ollman v. Evans*, 750 F.2d 970, 974–75 (D.C. Cir. 1984) (en banc) (same). After *Milkovich*, the First Circuit examined the issue closely and concluded that "*Milkovich* did not depart from the multi-factored analysis that had been employed for some time by lower courts seeking to distinguish between actionable fact and nonactionable opinion." Phantom Touring, Inc. v. Affiliated Publ'ns, 953 F.2d 724, 727 (1st Cir. 1992). Other lower courts have also employed a totality of circumstances approach. *See, e.g.*, Adelson v. Harris, 973 F. Supp. 2d 467, 488–89 (S.D.N.Y. 2013) (apply multifactor test); Yates v. Iowa West Racing Ass'n, 721 N.W.2d 762, 771 (Iowa 2006) (same); Wheeler v. Neb. State Bar Ass'n, 508 N.W.2d 917 (Neb. 1993) (same); Neumann v. Liles, 369 P.3d 1117 (Or. 2016) (same); *see also* Wampler v. Higgins, 752 N.E.2d 962 (Ohio 2001) (holding that Ohio's constitution required application of totality of circumstances approach even to nonmedia defendants); DAVID A. ELDER, DEFAMATION: A LAWYER'S GUIDE § 8:15 (2003) (noting that lower courts interpreted *Milkovich*'s protection of statements that do not imply assertion of actual facts in "widely varying ways").

199. Rodriguez v. Panayiotou, 314 F.3d 979, 986 (9th Cir. 2002); *see also* Obsidian Fin. Group, LLC v. Cox, 740 F.3d 1284, 1293 (9th Cir. 2014), *cert. denied*, 134 S. Ct. 2680 (2014); Gardner v. Martino, 563 F.3d 981, 986–87 (9th Cir. 2009); Partington v. Bugliosi, 56 F.3d 1147, 1152–53 (9th Cir. 1995); Knievel v. ESPN, 393 F.3d 1068, 1074–75 (9th Cir. 2005) (articulating "totality of the circumstances" test as examining (1) "the statement in its broad context, which includes the general tenor of the entire work, the subject of the statements, the setting, and the format of the work"; (2) "the specific context and content of the statements, analyzing the extent of figurative or hyperbolic language used and the reasonable expectations of the audience in that particular situation"; and (3) "whether the statement itself is sufficiently factual to be susceptible of being proved true or false"); Underwager v. Channel 9 Australia, 69 F.3d 361, 366 (9th Cir. 1995) (same).

setting, and the format of the work;

(2) the specific context and content of the statements, analyzing the extent of figurative or hyperbolic language used and the reasonable expectations of the audience in that particular situation; and

(3) whether the statement itself is sufficiently factual to be susceptible of being proved true or false.[200]

Using this broadly contextual "totality of circumstances" approach, courts have begun to take note of how social media contexts affect the interpretation of allegedly defamatory statements. For example, in *Feld v. Conway*,[201] a federal district court interpreted a defamatory tweet to the plaintiff saying "you are fucking crazy!"[202] to be mere hyperbole by looking at the Twitter comment thread of which it was a part. Looking at the "totality of circumstances,"[203] the court noted that a "tweet cannot be read in isolation, but in the context of the entire discussion."[204] The court even suggested that refusing to read the individual tweet literally, as an assertion of mental instability, but instead reading it in the comment-thread context was the "way in which a reasonable person would interpret it."[205] Thus, the court insisted that the evaluation of a defamatory statement must take account of its social media context.[206]

Drawing lessons from the intersection of tort law and the First Amendment, criminal law should not impose liability upon speech unless, when viewed in context and examining the totality of circumstances, it would have been *reasonably* perceived by the immediate parties to the communication as a threat.[207] The immediate parties would include the specific target, if any, and the immediate audience, if any. The line drawn here tilts the scale toward free expression by protecting speakers acting consistently with the norms of linguistic

200. *Rodriguez*, 314 F.3d at 986.

201. *See* 16 F. Supp. 3d 1 (D. Mass. 2014); *see also Obsidian*, 740 F.3d at 1284 (applying similar test in blog defamation case focusing on "(1) whether the general tenor of the entire work negates the impression that the defendant was asserting an objective fact, (2) whether the defendant used figurative or hyperbolic language that negates that impression, and (3) whether the statement in question is susceptible of being proved true or false").

202. *Feld*, 16 F. Supp. 3d at 2–4.

203. *Id*. at 3.

204. *Id*. at 4.

205. *Id*.

206. *Id.* For further examples, see Jones & Lidsky, *supra* note 115, at 159–72.

207. Some might contend, however, that threatening or violent speech is so antisocial and has such minimal value that any speech that a reasonable outside observer would view as threatening merits punishment, even if those immediately involved in the communication do not view it as a threat. After all, social media communications differ from oral communications precisely because it is likely they can be viewed by those unfamiliar with the immediate contexts, and the speaker must assume that risk (and the criminal liability that goes with it) because his misconstrued social media speech has the capacity to inflict social harm. *See* Eric J. Segall, *The Internet as a Game Changer: Reevaluating the True Threats Doctrine*, 44 TEX. TECH. L. REV. 183, 196 (2011) (arguing the Court should "make clear that personally directed attacks on individuals that could reasonably be interpreted as a real threat to that person are not protected by the First Amendment . . . .").

subcultures, even where outside observers unfamiliar with social media contexts might view their speech as threats. Tilting the balance in this manner gives adequate breathing space for the continued vitality of social media discourse.[208] Even more importantly, tilting the balance in this manner will prevent police and prosecutors from overreaching and singling out for punishment speakers who strike them as deviant.[209] And given that many of these speakers may be young, politically powerless, or expressing unpopular beliefs, the selective prosecution of hyperbolic speech misconstrued as threats raises significant First Amendment concerns.[210] Here, the dangers of sliding down the slippery slope to speech suppression are especially pronounced as a result of the negative social perception of the speakers.[211]

Incarcerating teenagers for hyperbole certainly sends them a message: one suspects that Justin Carter will never again use social media in the joking manner he did when he sent his alleged threat over Facebook. But if the goal is more civility in social media, criminalization seems like a poor tool for social change.[212] Moreover, the failure to draw precise lines between threats and hyperbole risks the overcriminalization of innocent speech and innocent speakers, many of whom are likely to be young people, simply because their speech is more likely to be misinterpreted by legal decision-makers unfamiliar with social media norms and conventions.

---

208. Whether the current uncertainty over questions of speaker intent and the role of context in true threats doctrine is likely to chill social media speakers is an unknowable empirical issue. Social media seem to generate an almost unstoppable flood of hyperbole that might be misconstrued as threats, and it is by no means clear that an increasing number of threats prosecutions will stanch the flow. Internet trolls may always be with us, regardless of whether criminal law cracks down on deeply uncivil behavior. *See generally* Buckels, Trapnell & Paulhus, *supra* note 125 (study of the psychology of Internet trolls).

209. If one thinks police or prosecutorial overreach is unlikely, consider the U.K. case known as the Twitter Joke Trial. In that case, *Chambers v. Director of Public Prosecutions*, [2012] EWHC 2157 (Admin) ¶¶ 12, 38, the High Court of England and Wales reversed a conviction of a man who had jokingly tweeted: "Crap! Robin Hood Airport is closed. You've got a week and a bit to get your shit together otherwise I am blowing the airport sky high!!" *Id.* ¶ 12. Police did not respond to the tweet until a week after it was made. *Id.* ¶¶ 13–14. The conviction was reversed on the basis the tweet "did not constitute or include a message of a menacing character." *Id.* ¶ 38. The court stated that language may not be construed as menacing "if the person or persons who receive or read it, or may reasonably be expected to receive, or read it, would brush it aside as a silly joke, or a joke in bad taste, or empty bombastic or ridiculous banter." *Id.* ¶ 30.

210. *See generally* Michal Buchhandler-Raphael, *Overcriminalizing Speech*, 36 CARDOZO L. REV. 1667 (2015) (discussing similar observations regarding the overcriminalization of speech).

211. *See* Frederick Schauer, *Slippery Slopes*, 99 HARV. L. REV. 361, 377 (1985) (explaining that the dangers a "decisionmaker's negative view of the parties is likely to lead to mistakes of a particular kind, to oversuppression rather than undersuppression, in the application of free speech principles, and these mistakes serve to create the special slippery slope danger").

212. *See* Aya Gruber, *Rape, Feminism, and the War on Crime*, 84 WASH. L. REV. 581 (2009) (discussing the shortcomings of criminal sanctions in fostering social change).

IV.
CHALLENGES FOR LEGAL DECISION-MAKERS AND A PROPOSAL FOR OVERCOMING THEM

Understanding the context of social media comments provides challenges for legal decision-makers. Serving judges, practicing attorneys, and jurors, especially those past the first bloom of youth, may not be familiar with the architectural features of different social media platforms, conventions of discourse within each social medium, or norms of discourse governing online gaming forums or other subcultures. After reading accounts of youths cursing at each other in broken English, punctuating with emojis or gifs, and typing or speaking while in the throes of mowing down electronic fiends, some decision-makers might even feel that jail time with access to nothing but a paper library might do them some good. Yet, justice systems should never deploy criminal sanctions lightly, especially when those sanctions target crimes committed solely through speech. If the law is to avoid punishing speakers merely for failing to foresee that their words might be misunderstood by outsiders to their online conversations, the true threats doctrine must be modified. One modification, of course, is that the Supreme Court could interpret the First Amendment to require speakers to have a culpable mental state of purpose to make a threat or knowledge of the threatening character of their speech before the speech could be deemed a criminal threat. It would also help safeguard speech if the Court stated clearly that the First Amendment bars punishment of speech as a threat unless that speech could reasonably be perceived as such. This clarification would solve the circuit split left unresolved by *Elonis*. But even with this clarification, juries and judges can infer purpose or knowledge from circumstances that they interpret incorrectly because they do not understand social media contexts. Put simply, law enforcement, prosecutors, judges, and juries do not know what they do not know about the interpretation of social media speech, and the defendant's own words are likely to be the most compelling evidence of intent and the most compelling evidence of whether a threat has occurred. Thus, merely clarifying existing law in the ways described above will not adequately prevent misinterpretations resulting in criminal liability. Therefore, this Section focuses on ways to allow full consideration of context in threats cases, especially those dealing with social media threats.

The first way to bolster legal decision-makers' understandings of social media speech alleged to be criminal is by the introduction of expert witnesses. Expert witnesses could offer a way to explain social media architecture and other unfamiliar elements of social media discourse to the legal community.[213] Much

213. Arguing for admission of expert testimony regarding the interpretation of social media speech is consistent with the admission of expert witness testimony concerning the interpretation of "drug slang" or "gang jargon," which routinely is allowed in criminal cases. *See, e.g.*, United States v. Akins, 746 F.3d 590, 598 (5th Cir. 2014) (holding that district court did not abuse discretion in admitting testimony regarding interpretation of "drug jargon"); United States v. Griffith, 118 F.3d 318, 321 (5th

as one might hire a doctor to put into layman's terms a baffling array of scientific terms, one could hire an expert to explain why the relevant context might recast speech that initially appears to be a true threat as protected speech. Attorneys conversant with the conventions and characteristics of social media discourse might also teach or sponsor continuing education courses to help other attorneys understand the relevant contexts of alleged threats.

Legal decision-makers, though, would benefit more directly from clearer guidelines from appellate courts directing them how to interpret alleged threats made in social media. As suggested by the discussion of defamation law above, First Amendment doctrine setting the boundaries of true threats should guide legal decision-makers to interpret threats based on the totality of surrounding circumstances. Such factors might include the following:

(1) *the architecture of the social media platform, including space limitations on individual posts, if any*. The architecture of a platform may fundamentally affect the sender's message. In Twitter, for example, the 280-character limit may mean that the speaker must use abbreviations, hashtags, or emojis that may be misunderstood, or the speaker's meaning may only become clear when a string of tweets are read in context. Likewise, the fact that Snapchat messages are designed to disappear soon after being read might make a threat seem more or less serious, depending on the surrounding contextual factors, such as the relationship between the parties. Nonetheless, the architecture of different social media may affect both the intended and received message.

(2) *the actual language used, including whether the alleged threat was conditional, specific, or threatened immediate harm and whether the speaker used emojis, hashtags, gifs, abbreviations, or profanity*. The actual language used is already considered in threats cases. However, social media threats cases are especially likely to contain language that is not familiar to the cross section of the public from which juries are drawn. In the Justin Carter case, for example, neither the prosecutor nor the judge seemed to be impressed with the argument that he had followed his seeming threat with "J/K" to indicate that he was just kidding. Perhaps they did not know the meaning of the abbreviation, or perhaps they were simply unconvinced that it altered either his intent or the effect of the post. Nonetheless, the specific language a speaker uses obviously affects meaning

---

Cir. 1997) ("Drug traffickers' jargon is a specialized body of knowledge, familiar only to those wise in the ways of the drug trade, and therefore a fit subject for expert testimony."); People v. Anderson, 149 A.D.3d 1407, 1413 (3d Dep't 2017) (allowing expert testimony regarding drug slang). In one such case involving gang jargon, the court noted that the proffered expert testimony was "based on standard principles of language interpretation, the foundation of which is repetition and context." *See* United States v. Diaz, 2006 WL 2699042, *3 (N.D. Cal. 2006).

and that language must be interpreted correctly if the speaker's meaning is to be discerned.

(3) *the overall tone of both the specific post and the conversation, if any, of which it is a part*. Just like informal conversations offline, conversations online can range from highly formal and serious to highly informal and facetious, sarcastic, bantering, or hyperbolic. Where a post or conversation falls along this range helps determine its meaning. "I'll kill you" when said to an estranged ex-wife as part of a broader unhinged tirade, as in the *Elonis* case, should be interpreted differently than a wife posting on Facebook "I'll kill you ☺" in response to her husband's post taunting her that he ate the last piece of cheesecake in the refrigerator at home. Of course, we know this intuitively, but it is helpful to bring this intuition to the fore when dealing with social media contexts with which not everyone is familiar.

(4) *the likely characteristics of the immediate audience, including the likely size and demographics of the audience and the relationship between the speaker and his or her audience*. One speaks to friends and lovers differently than one speaks to strangers, and young people tend to speak to each other differently when parents are not listening than when they are. These contextual factors matter when interpreting true threats, just as they matter when interpreting defamation. The age, education, sex, and race of the speaker and his or her audience also affect interpretation.[214] For example, the Supreme Court in *Watts* took notice that the audience of the teenage speaker's purported threat to shoot then-President Lyndon Johnson was also young and also opposed to the Vietnam War, as well as the fact that they laughed upon hearing the "threat." In *Watts*, these contextual factors were obvious to the Court, but the Court would be far less likely to understand contextual features that would be similarly obvious to social media "insiders," which is why explicit examination and expert testimony may sometimes be necessary.

(5) *the social media platform used and the conventions of speaking within that platform, and whether the speech occurred within a distinctive subforum*. Some social media platforms, put simply, host more restrained discourse than others. Certain subforums

214. *See, e.g.*, Sherron B. Kenton, *Speaker Credibility in Persuasive Business Communication: A Model Which Explains Gender Differences*, 26 INT'L J. OF BUS. COMM. 143, 143 (1989) (surveying research on source credibility and gender differences); Lawrence Hosman, *Powerful and Powerless Speech Styles and Their Relationship to Perceived Dominance and Control*, *in* THE EXERCISE OF POWER IN COMMUNICATION 221, 221 (R. Schulze et al. eds,, 2015) ("[R]eport[ing] the findings of over 30 years of research on powerful and powerless speech styles and their relationship to dominance and control.").

on Reddit, for example, are replete with violent rhetoric and hate speech, which would, in some instances, lend credence to an alleged threat and in others mark it as hyperbole when blended with other factors.

(6) *the surrounding context of the speaker's posting, including whether it was a response to another speaker's post*. If a teen girl posts a photo on Facebook and her friend responds "Dead!" and then they carry on the conversation about where to go to lunch, the context tends to suggest the word "dead" was not a threat. Taken out of context, however, the possibility for misunderstanding is high. At a minimum, therefore, legal decision-makers should consider alleged threats set in their immediate surrounding contexts in order to avoid the risk of criminalizing innocent speakers.

Even equipped with a list of contextual factors to consider, however, police, prosecutors, judges, and juries may still struggle to correctly interpret social media speech. Thus, we propose the creation of a more formal "context defense" to make sure the First Amendment rights of those accused of making criminal threats are safeguarded.

The defense would work as follows. Once the accused is charged with making a criminal threat, the accused may invoke context as a defense to liability. Once the accused invokes the defense, the trial court judge should conduct a pretrial hearing to establish two prerequisites to proceeding to trial. First, the judge should evaluate whether there is probable cause to believe that the defendant actually made the alleged threat. This is a concern in online settings where digital images or communications may be altered,[215] and the Internet includes so many avatars and aliases that it may not always be clear that the person accused of making threats actually made the threats. The burden of proof on this issue should be on the prosecution because attaching wrongful conduct to the actual defendant is an essential prerequisite to imposing criminal liability.

If this threshold is met, the court should give the defendant an opportunity to show that the statement made or posted was not an actual threat based on the totality of the surrounding context. In examining context, it is important to understand that context affects both the mens rea and actus reus of the offense. Context can be circumstantial evidence that the defendant possessed the requisite mens rea to make a threat, or, conversely, context can negate the requisite mens rea. By the same token, the actus reus of the threat exists by virtue of the effect

215. Online impersonation is a significant enough problem that a number of states have adopted laws criminalizing it. *See, e.g.*, CAL. PENAL CODE § 528.5 (West 2013); HAW. REV. STAT. § 711-1106.6 (2008); MISS. CODE. ANN. § 97-45-33 (West 2011); N.Y. PENAL LAW § 190.25(4) (McKinney 2008); TEX. PENAL CODE ANN. § 33.07 (West 2011).

of the defendant's words on his target and/or his audience, and context affects how reasonable targets or audiences perceive the defendant's words.[216]

At the pretrial hearing, the defendant should have a right to bring an expert witness or otherwise present probative evidence that his statements were not threats when viewed in their full context. If the court determines that there is probable cause to believe that the defendant *intended* (that is, either purposely or knowingly) to make what a reasonable target of the threat or reasonable reader within the immediate audience of the post would view as a true threat,[217] then the case may proceed to trial. If the court determines that, when viewed in the appropriate context, the defendant's alleged threats were not true threats, then the case should be dismissed.[218]

This context defense procedure is not a radical departure from existing doctrine. Such pretrial hearing procedures are used in other contexts[219] to safeguard important rights such as the Sixth Amendment right to assistance of counsel[220] or the Fourth Amendment right to be free of illegal searches and seizures.[221] Thus, it is logical that such pretrial hearings should be used to protect criminal defendants' First Amendment rights as well.

This brief hearing process has the benefits of moving the matter off of a judge's docket more quickly than a trial, saving the state the resources of prosecuting a defendant for uttering potentially protected speech, and sparing an innocent person the costs of mounting a full defense, not to mention sparing him or her the unfortunate consequences of extended incarceration. The hearing

---

216. This so-called defense is what is sometimes called a "failure-of-proof" defense, that is, "one in which the defendant introduces evidence . . . demonstrat[ing] that the prosecution has failed to prove an essential element of the offense charged." *See* JOSHUA DRESSLER, UNDERSTANDING CRIMINAL LAW 204 (6th ed. 2012).

217. It may sometimes be difficult to ascertain who the "immediate audience" of a social media post is, but consideration should be given to factors such as whether the defendant limited his or her postings to "friends," that is, to a selected group; whether he or she addressed or targeted a particular person or groups of persons; and whether the post was part of an ongoing dialogue with a particular person or groups of persons. If, as in the Justin Carter case, the defendant's posting was a direct response to another speaker's posting, that speaker should be considered the immediate audience, as well as any others who have commented on the dispute contemporaneously.

218. The burden of establishing probable cause remains on the prosecution.

219. Probably the most familiar example is the pretrial suppression hearing, in which courts are asked to determine which evidence is inadmissible in the defendant's criminal trial. *See* Brewer v. Williams, 430 U.S. 387, 406 (1977) (confession obtained in violation of defendant's right to advice of counsel); Miranda v. Arizona, 384 U.S. 436, 478–79 (1966) (evidence obtained through interrogation absent waiver of fifth amendment rights); Jackson v. Denno, 378 U.S. 368, 376–77 (1964) (involuntary confession); Mapp v. Ohio, 367 U.S. 643, 655 (1961) (evidence seized in violation of fourth amendment); Nardone v. United States, 308 U.S. 338, 341–42 (1939) (evidence obtained as a result of knowledge gained from illegal wiretap); Silverthorne Lumber Co. v. United States, 251 U.S. 385, 392 (1920) (evidence seized as an indirect result of fourth amendment violation); Weeks v. United States, 232 U.S. 383, 398 (1914) (evidence seized in violation of fourth amendment in federal trial).

220. *See Brewer*, 430 U.S. at 406 (holding inadmissible confession obtained in violation of defendant's right to advice of counsel).

221. *See Mapp*, 367 U.S. at 655 (holding inadmissible evidence seized in violation of the Fourth Amendment).

process might also curtail any tendencies of prosecutors to overreach and charge dim youths with crimes such as making terroristic threats, when they have done nothing more than to make hyperbolic and ill-advised comments in a public forum while playing a computer game.

For those defendants who actually go to trial, a defense based on context should be available. Asserting the context defense would allow the defendant to bring expert testimony to help explain the proper interpretation of his or her words. That defense would allow a defendant to show that although the defendant made certain statements, those statements were, when viewed within the proper context, not true threats because the defendant lacked the requisite intent to threaten.

Some might contend that since the burden of proof is already on the prosecution to establish the actus reus and mens rea of the offense the context defense is unnecessary. Yet prosecutors, judges, and juries do not know what they do not know. They may assume they share interpretive conventions with the defendant that they do not in fact share, especially in cases involving threats made in social media by members of a younger generation. The context defense therefore provides a mechanism to supply or enrich the relevant context and focus attention on the kinds of factors that should ultimately determine guilt or innocence in social media threats cases.

As another way of focusing decision-makers' attention in the right place, invoking the context defense should entitle the criminal defendant to the benefit of a special instruction explaining the contextual factors that may be relevant in deciding whether the defendant's words were threats or mere hyperbole. After specifying the relevant factors, the instruction might clarify that the burden of proof is on the prosecution and simply ask the jury to determine whether the defendant committed the offense beyond a reasonable doubt based on the totality of circumstances. The more focused inquiry supplied by the special instruction would then lend itself to more informed judicial review.

CONCLUSION

This Article, put simply, represents a plea for more precision in drawing the line between trash talk and terrorist threats. It may seem counterintuitive to argue for the protection of those whose loose language can be easily misunderstood as a threat. Yet First Amendment theory's concerns about selective targeting of disfavored speakers and its demand that speech cause cognizable harm in order to be actionable weigh against criminalizing speech that was not intended to threaten and/or was not understood by its targets as intimidation or threats. Overcriminalization concerns should be especially acute in cases in which defendants are disproportionately likely to be adolescents engaged in discourse over social media. This is true even though the prevalence of violent and hateful speech in social media is a formidable concern for those of us committed to fostering meaningful and inclusive public discourse.

Nonetheless, it is important to ensure that the proffered cure for disordered discourse in social media is not worse than the disease. This Article has shown that the interpretation of social media speech by legal decision-makers is rife with possibilities for misunderstanding and has tendered an antidote. This antidote, in the form of a context defense, allows those charged with making terroristic threats to show, at a very early stage of criminal proceedings, the context in which their alleged threats occurred negated their criminal culpability. Moreover, the defense provides a guide for legal decision-makers at trial as to the factors that influence proper interpretation of alleged threats within social media. This, and no less, is what courts should interpret the First Amendment to require, just as they have done in other legal contexts.

Figure 1: A screenshot of the post by Justin Carter that gave rise to his prosecution.[222]

222. Hollingsworth, *supra* note 8.