To: SEALS Works-in-Progress Panel
From: Aaron Caplan
Re: Fight-or-Flighting Words
Date: August 1, 2018

I look forward to exchanging ideas with you next week.

My current draft of this work in progress has some topics pursued at excessive length and obsessive detail, while others receive quite scanty treatment. Circulating that particular document at this stage would be unkind to readers!

For our discussion purposes, I boiled things down into the accompanying document that may best be described as an expanded outline. Most quotations and citations are omitted, and the prose will be uninspiring. The goal is to create a document that is at least evenly sketchy from start to finish. It will indicate the structure and primary arguments of the piece better than the current draft does, and be less painful of a read.

# Fight-Or-Flighting Words

To: SEALS Works-in-Progress Panel
From: Aaron Caplan
Re: Fight-or-Flighting Words
Date: August 1, 2018

# INTRODUCTION

When may the government punish insults (defined as a person-to-person expression, in whatever form, of the idea "I do not respect you")? Of the possible reasons for proscribing insults, only one has been endorsed by the US Supreme Court as sufficient grounds to justify an exception to the First Amendment's general rule protecting speech. Specifically, insults may be punished only if they are "fighting words," that is, insults that will cause the insulted person to respond with retaliatory violence. The image called to mind is of a card game in a saloon in the Old West, where one cowboy calls another a lily-livered four-flusher, which causes the insulted cowboy to declare "them's fighting words!" before unleashing (presumably well-deserved) violence on the insulter.

The cases establishing that fighting words are a proscribable category of speech date to the early 1940s, most prominently *Chaplinsky v. New Hampshire* (1942). The precise legal description of fighting words has varied, but a frequent definition is found in *Cohen v. California* (1971): "those personally abusive epithets which, when addressed to the ordinary citizen, are, as a matter of common knowledge, inherently likely to provoke violent reaction." This formula assumes a particular set of beliefs about human psychology. It assumes that when insulted in just the right way, a person will enter what amounts to a primordial fight-or-flight response beyond conscious control, where emotions overwhelm reason, leading to an unacceptably high likelihood that the person will respond by fighting. To avoid this potential for violence, the law punishes the insulter who places the insulted person in such a mental state.

There are relatively few statutes specifically making it unlawful to utter insults that will provoke retaliatory violence from the insulted person. Instead, the fighting words doctrine is invoked when examining convictions under such broadly-defined charges as breach of the peace, disorderly conduct, or use of abusive language. In significant part, the 20th-century fighting words doctrine originated as an effort to cure the potential vagueness or overbreadth that could arise when such nebulous charges were based solely on a defendant's speech.

In theory, the fighting words doctrine could be used to punish anyone who directs the right sort of insult to anyone else. In practice, it is overwhelmingly used to punish those who insult police officers. Loudmouths who insult police may find themselves arrested and (only occasionally) charged for offenses colloquially known as "contempt of cop." To deal with this problem, courts in some but not all jurisdictions have added a gloss to the fighting words doctrine, namely that since police officers ought not respond violently to insults, those who insult them should also not be convicted of fighting words offenses. This is in tension with the underlying premise behind the fighting words doctrine, namely that virtually all people will respond violently to certain insults. If

police officers can control their impulse to retaliate violently, why can't the ordinary person?

The fighting words doctrine has been criticized before on various grounds. This Article examines it in light of recent advances in psychology and neuroscience. Consider this article to be in the spirit of "behavioral realism," which encourages the law to reflect the realities of human behavior rather than inaccurate idealized models of it. The Article makes two primary assertions about human behavior in response to insults.

- First, the model of unconscious processes inevitably tending toward violence is seriously flawed. Sometimes someone may respond violently to an insult and sometimes not; the mechanisms contributing to the violent response are far more complex and context-specific than the folk psychology underlying the fighting words doctrine would admit.
- Second, if anyone is likely to respond violently to an insult, it will be a person seeking to defend a higher position in a social hierarchy against a challenge to status from someone located adjacent to or below them in that hierarchy. This means that police are far more likely than most people to engage in retaliatory aggression in response to insult.

In light of these understandings, the Article proposes that our free speech law should abandon retaliatory violence as a reason to proscribe insults.

Part I of the Article describes the fighting words doctrine as it currently exists. It begins by tracing the history of the doctrine, which has a great deal to do with the related problems of vagueness and policing. It next examines how current law deals with the frequent reality that the fighting words doctrine is used to justify arrest or prosecution of those who insult the police. As part of this survey, the Article will describe a data set I collected consisting of all state and federal fighting words cases decided in 2013 and 2014. The cases reveal some interesting patterns. (a) [[Approximately two-thirds]] of the cases that mention the fighting words doctrine do so purely in dicta, where the fighting words category is offered as a convenient example of a proscribable category. (b) Of the cases where the fighting words doctrine is necessary for the outcome, [[approximately 80%]] involve insults directed to police or private security officers, and the remainder tend to involve insults to other authority figures, usually governmental (e.g., wildlife inspectors). (c) Of the litigated cases involving insults to police, [[only a minority]] were ones where prosecutors saw fit to bring charges. These cases came to court as a result of a civil suit by the arrested-but-not-charged insulter.

Part II of the Article describes the neuroscience of responding to insults, contrasting it to the folks psychology underlying the fighting words doctrine. The main points in this section include the following. (a) The fighting words doctrine presumes a sharp dichotomy between reason and emotion; and that when strong emotions are triggered, reason (and the ability to refrain from violence) is helpless. While there is a grain of truth in this depiction – sometimes strong emotions can make it difficult to exercise some cognitive functions – the picture is seriously incomplete. Reason and emotion work hand in hand in most settings, including when one is insulted. (b) The

tendency to respond violently to insults is a hallmark of what sociologists call the culture of honor. Cultures of honor tend to arise in settings where there is no effective rule of law, and where people's best protection for their resources is to develop a reputation for fiercely protecting both their assets and their status. Hence the popular linkage between the fighting words doctrine and frontier societies like the Old West. Responding to insults with violence is not a biological imperative (even though people who have thoroughly assimilated the values of a culture of honor may respond biologically in characteristic ways when insulted). (c) Whether a person under threat will respond by fleeing or fighting depends on a huge array of context-specific factors. But as a general matter, people within a social hierarchy will direct their aggression at subordinates. This suggests that insults directed from the top down will result in freezing or fleeing, not fighting. Insults like these do not challenge positions within the hierarchy; they are part and parcel of life within it. By contrast, violence is most likely to occur when insults threaten status within the hierarchy, which occurs when the insults are directed upwards or laterally by rivals for social status.

Part III of the Article considers how the law should accommodate these insights. At this stage, I am proposing three lessons for the law.

(1) The prospect of retaliatory violence should not be relied upon as a reason to make certain insults unlawful. (a) As a general matter, our law should not be premised on faulty psychological models. (b) The culture of honor makes sense only in the absence of the rule of law. Indeed, the culture's internal logic involves self-help (a real man deals with an insult with his fists, not by invoking the law). Our law ought not model itself after the vigilantism of a frontier society. (c) The fighting words doctrine has the effect of privileging existing status hierarchies, because the words made illegal are those of subordinates challenging the status of superiors. This makes for a dubious doctrine in an open society. (d) *Chaplinsky* relied on the notion of retaliatory violence to solve a vagueness problem, but that effort cannot succeed. Given how varied the contextual factors are that could lead to retaliatory violence, using that as our yardstick for which words are proscribable will inevitably be unpredictable. This means the doctrine does not help avoid the twin evils that the vagueness doctrine seeks to eliminate: lack of notice to the citizen and the opportunity for arbitrary power by law enforcement.

(2) Unlike the fighting words doctrine, the existing true threats doctrine is a good fit with the neuroscience described in Part II. When someone is threatened with violence, a fight-or-flight response may well result. The harms associated with true threats (insecurity, fear, and withdrawal from discourse) are consistent with how people behave on the freeze/flee end of the fight-or-flight response. This observation helps explain why the fighting words justification for a law against cross burning was unpersuasive in *RAV v. St. Paul* (1993), while a true threats justification for a law against cross burning was intuitively and legally successful in *Virginia v. Black* (2000).

(3) Might there be reasons for proscribing insults that do not rely on the notion of retaliatory violence? While this Article does not take any position on the question, it may be possible to argue that the psychic injury of an insult – especially when delivered from

the top down – may be a suitable basis for legal remedies. If justifiable, it would represent a broadening of the true threat doctrine, not of the fighting words doctrine.

# I. The Fighting Words Doctrine in Theory and Practice

Our commitment to free speech means that speech may not be banned solely because it is disliked, unwanted, or discomfiting – all terms that would apply to an insult. Moreover, one person's insult can be another person's term of endearment. This means that a law flatly banning all insults would be unconstitutionally vague, overbroad, or both. Recognizing this, current First Amendment law says that only some insults are proscribable.

But on what basis do we identify the proscribable insults? Possible reasons might be that the insults are offensive or distasteful to the insulted person or those who overhear the insult; that the insult will lead to breach of the peace in the form of retaliatory violence (against the insulter by the insultee); that the insult incites others to imminent violence (likely against the insultee or those in a similar group); or that the insult conveys a true threat. Making it more difficult is that these categories can and do overlap in practice. Depending on the context, one person telling another to "go to hell" might be part of a true threat, an offense to sensibilities, or a prompt to retaliate violently. Of these possible explanations, existing free speech law renders one of them an improper basis to proscribe speech. The fact that speech is offensive, upsetting or distasteful to listeners' sensibilities has been squarely rejected as a basis for speech regulation. See *Terminiello v. Chicago* (1949), *Street v. New York* (1969), *Bachellar v. Maryland* (1970), *Coates v. Cincinnati* (1971). Whether an insult forms part of a proscribable incitement under *Brandenburg v. Ohio* (1969) or true threat under *Watts v. US* (1969) will be judged under the tests applicable to those doctrines. This leaves the stated premise of the fighting words doctrine: retaliatory violence. That concept is supposed to answer two questions: (a) which insults may be proscribed? And (b) why proscribe them?

## A. The Fundamentals

The basics of the fighting words doctrine can be seen in three opinions.

### 1. *Cantwell*

Although it is not the part of the opinion most often cited today, part of *Cantwell v. Connecticut* (1940) involved a conviction for the common law offense of inciting a breach of the peace. Jesse Cantwell, a proselytizing Jehovah's Witness, asked two men on the street if he could play them a recording. They agreed; the recording attacked Roman Catholicism; the two men, both Catholics, later testified that they thought about responding violently (although they did not). The Supreme Court held that Cantwell had not behaved in a discourteous way; he uttered no insults and hence there was insufficient evidence that he could have incited a breach of the peace.

Along the way, the Court said: "Resort to epithets or personal abuse is not in any proper sense communication of information or opinion safeguarded by the Constitution, and its punishment as a criminal act would raise no question under that instrument." In saying that insults were not protected, the Court did not limit its scope to those insults that would prompt a retaliatory response. Nor did the Court explain why insults were not protected.

### 2. *Chaplinsky*

*Chaplinsky v. New Hampshire* (1942) is well remembered as the origin of the fighting words doctrine, even though it quotes and cites *Cantwell*. Walter Chaplinsky was a Jehovah's Witness who was threatened and then assaulted by a hostile mob. The local police refused to arrest any of those who attacked Chaplinsky, but they did take him into protective custody. When he asked the marshal why no one was arresting the ones who started the fight, he was told: "shut up you damn bastard and come along." At this point, Chaplinsky told the marshal: "you are a damned fascist and a racketeer." The marshal responded by calling Chaplinsky "an unpatriotic dog," and another arresting officer said "you son of a bitch, we ought to have left you to that crowd there and [let] them kill you." The New Hampshire Supreme Court later observed drily: "It may be remarked that nobody concerned, taking Chaplinsky at his word, used proper restraint on this occasion."

Chaplinsky was convicted of violating a statute that forbade addressing "offensive, derisive, or annoying words" to others in public places. In response to a vagueness challenge, the New Hampshire Supreme Court held that despite appearances, the statute was limited to "words likely to cause an average addressee to fight." The US Supreme Court held that the statute was not vague as so limited. Chaplinsky "need not therefore have been a prophet to understand what the statute condemned." As for the precise words spoken, the Court was willing to take judicial notice "that the appelations 'damn racketeer' and 'damn Fascist' are epithets likely to provoke the average person to retaliation, and thereby cause a breach of the peace."

With the vagueness problem resolved, the Court had no trouble finding that insults were constitutionally proscribable, in a paragraph widely quoted ever since.

> There are certain well-defined and narrowly limited classes of speech, the prevention and punishment of which have never been thought to raise any Constitutional problem. These include the lewd and obscene, the profane, the libelous, and *the insulting or 'fighting' words—those which by their very utterance inflict injury or tend to incite an immediate breach of the peace.* It has been well observed that such utterances are no essential part of any exposition of ideas, and are of such slight social value as a step to truth that any benefit that may be derived from them is clearly outweighed by the social interest in order and morality.

The italicized language is often quoted as a definition of fighting words, but the Court's opinion did not rely on the concept that insults "by their very utterance inflict

injury." The decision rested solely on the notion of retaliatory violence. Its logic was not too far removed from the incitement doctrine of *Schenck* (1919): Since the government has power to punish breaches of the peace, it should also be able to punish speech that elicits it.

However, neither *Chaplinsky* nor any other majority opinion of the Supreme Court has attempted to explain why an insult would provoke retaliatory violence. The closest any opinion of the Court has come to this is Justice Jackson's dissent in *Kunz v. New York* (1951).

### 3. The *Kunz* Dissent

The City of New York denied to Carl Jacob Kunz, a Baptist minister, a permit to hold public worship meetings on the street. The denial seems to have been based on the fact that Kunz on previous occasions had included insulting messages in his sermons, such as "the Pope is the anti-Christ" and that Jews are "Christ-killers." The majority of the Court held that it was an unconstitutional prior restraint to deny a permit under an ordinance that gave city official unfettered discretion to decide whether to grant or deny it. (This was consistent with a similar ruling in another part of *Cantwell*.)

Justice Jackson – fresh from his experience as a prosecutor in the Nuremberg trials – dissented, much as he had dissented in the factually analogous *Terminiello v. Chicago* (1949) (city arrests volatile anti-Semitic speaker). The insults used by Kunz were "more clearly fighting words" than the epithets "racketeer" and "Fascist" from *Chaplinsky*.

Since Kunz was holding his meetings on the street, the question for Justice Jackson was whether a city "must place its streets at his service to hurl insults at the passerby."

> These terse epithets come down to our generation weighted with hatreds accumulated through centuries of bloodshed. They are recognized words of art in the profession of defamation. They are not the kind of insult that men bandy and laugh off when the spirits are high and the flagons are low. They are not in that class of epithets whose literal sting will be drawn if the speaker smiles when he uses them. They are always, and in every context, insults which do not spring from reason and can be answered by none.

The insulted person reacts violently just as predictably as a theater-goer will react with panic when hearing shouts of fire. This is a universal reaction shared by all, and regardless of context. (Justice Jackson could not have imagined the comedy of Richard Pryor.)

## B. The Police Gloss

It is not a new observation that police may use nebulous penal statutes to arrest people they consider to be a thorn in the side, even if they have not committed crimes. Indeed, one vice of a vague statute (like the unfettered discretion in the permitting systems in *Cantwell* and *Kunz*) is that it invites arbitrary enforcement. The drafters of the Model Penal Code in 1962 included commentary in their proposed disorderly conduct statute, §250.2, indicating that use of disorderly conduct statutes against those who criticize or argue with police is a serious constitutional problem, and suggesting that adopting jurisdictions could include a proviso that speech directed to police is beyond the reach of the statute.

The concern over police reliance on the fighting words doctrine (and its associated statutes) has motivated some unease in SCOTUS opinions. Case after case involve facts like *Chaplinsky*, where the alleged fighting words were uttered to police by someone who rightly or wrongly feels ill-treated by them. There has been no categorical announcement that the constitution protects insults against the police. However, SCOTUS has often used related doctrines of vagueness and overbreadth to void the statutes used to justify arrest or punishment of those who insult police.

In the early 1970s, SCOTUS considered a string of cases that combined concerns over (a) police misuse of fighting words statutes, especially against antiwar or civil rights demonstrators; with (b) the contemporaneous concern about constitutional protection for vulgar or profane words, as in *Cohen v. California* (1971) ("fuck the draft") and *Papish v. University of Missouri* (1973) (underground newspaper using the word "motherfucker"). The usual result in these cases was to find statutes overbroad because they criminalized more than just fighting words as defined in *Chaplinsky*.

For example, the defendant antiwar protestor in *Gooding v. Wilson* (1971) said to arresting officers things like "You son of a bitch, I'll choke you to death" – a combination of an insult and a threat. A majority of the Supreme Court overturned the conviction because it was premised on a statute forbidding "opprobrious words or abusive language" that had allowed past convictions for mere utterance of profanity or insults deemed too mild to be fighting words. In *Lewis v. New Orleans II* (1974), the defendant called officers "you god damn motherfucking police" and was convicted under a local ordinance forbidding "to curse or revile or to use obscene or opprobrious language toward" city police. SCOTUS invalidated the statute as overbroad. In a concurrence, Justice Powell suggested that "a properly trained officer may reasonably be expected to exercise a higher degree of restraint than the average citizen, and thus be less likely to respond belligerently to fighting words."

The Supreme Court revisited the issue in *Houston v. Hill* (1987), where a gay rights activist observing an arrest told police "why don't you pick on someone your own size" – a phrase that depending on context could be an attempt to induce shame or an invitation to a brawl. Police arrested him FOR interrupting an officer in the discharge of duties. SCOTUS found the statute overbroad, because it could be used to criminalize mere criticism of police. "The freedom of individuals verbally to oppose or challenge

police action without thereby risking arrest is one of the principal characteristics by which we distinguish a free nation from a police state." Did this mean that police could never arrest for insults? No, said the Court in a footnote that the right to vergbgally challenge police would not encompass the utterance of fighting words (whatever those might still be).

## C. Fighting Words Today

With a research assistant I coded all opinions from federal or state courts available on Westlaw that included the phrase "fighting words." The numerical tally of the approximately 130 cases has not yet been completed, but here are some quick and dirty take-aways from the data set.

*Mere Mentions*. [[At least 2/3]] of the cases in the data set were mere mentions. Just as an offhand reference to "the lewd and obscene" and "the profane" in *Chaplinsky* was used as a gesture toward the general idea of proscribable categories, "fighting words" is such a construct today.

*Fighting Words Directed at Police Officers*. Of the cases where a Court had to grapple with the fighting words doctrine, [[approximately 80%]] involved a person who allegedly directed fighting words to a police officer or private security officer. The remainder tended to be various other authority figures, usually governmental (e.g., wildlife inspectors). Only [[a handful]] involved words directed to other civilians.

*Arrests v. Prosecutions*. Of the police cases, [[under half]] were actual criminal prosecutions. The remainder were civil suits, where a person arrested or threatened with arrest but not prosecuted sought damages (typically under §1983) for wrongful arrest or violation of free speech rights, as in *Houston v. Hill*. That so many of these arrests were deemed not worthy of prosecution suggests that fighting words crimes can readily be used in spurious "contempt of cop" scenarios.

Why does the case law reveal such a prominent amount of legal action taken in response to insults to police? They cannot be the only people being insulted. Part of the answer may lie in resource allocation decisions within the criminal justice system. If a person calls 911 to report that their neighbor called them a hurtful epithet, it is doubtful that police or prosecutors would prioritize legal response. Moreover, in jurisdictions that follow the common law rule for arrests, police can only make a warrantless arrest for a misdemeanor if it is committed in the officer's presence. As a result, only insults to police may enter the legal system at all. But as described below, I believe there is more to it than that.

## II. Brain Science[1]

The folk psychology of the fighting words doctrine seems to include the following concepts:

A. The realms of reason and emotion are separate from each other. (In the ideal world, reason will always prevail, so the First Amendment does not protect speech that threatens reasoned discourse.)
B. When insulted, the emotions overwhelm reason. The result is an instinctual excited state that is highly prone to result in violence. (It is also a state where reasoned discourse becomes impossible.)
C. The tendency to respond to insults with retaliatory violence is universal and not context-specific.

Like most folks psychology, this framework is built on grains of truth. Reason and emotion can feel subjectively different and can involve different brain regions and chemistry; when excited by perceived threats, emotional centers may partially inhibit cognitive centers. But overall, the folk psychology of the fighting words doctrine is so incomplete as to be misleading if taken as a description of, or predictor of, human reactions to insults.

For this section, I take the fighting words doctrine at face value insofar as it purports to be about reactive violence or defensive aggression. Different types of aggression can arise in different circumstances. Indeed, defensive aggression (as opposed to, say, the aggression of a predator attacking prey or the aggression of rivals for mates within a species), tends to involve different brain circuitry. Cross-species aggression tends to be studied entirely different from within-species (i.e. conspecific) aggression. For our purposes, retaliatory violence is that which arises from a perceived threat or danger to the individual. The following sections consider what occurs when that danger comes in the form of the message that a speaker does not respect you. (I do not consider here, just as the law does not consider, what motivates the speaker to engage in the aggressive activity of delivering an insult.)

### A. The Interaction of Cognition and Emotion

From ancient Greece to Rene Descartes to the Romantic poets, folk psychology has contrasted the mind and the body; thinking and feeling; cognition and emotion. The latter is responsible and admirable, the latter hot-tempered and untrustworthy. This

[1] I'm neither a neuroscientist nor a sociologist. Most of what I describe here is drawn from one or more of these. Robert Sapolsky, *Behave: The Biology of Humans at Our Best and Worst* (2017); Joseph LeDoux, *Anxious: Using the Brain to Understand and Treat Fear and Anxiety (*2016); Elizabeth Johnston & Leah Olson, *The Feeling Brain: The Biology and Psychology of Emotions* (2015); Steven Pinker, *How the Mind Works* (2009); Jaak Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions* (2004).

dualistic view is out of favor. Cognition and emotion in fact work closely with each other; without emotional ability the results of our thoughts are not trustworthy.

Rejecting dualism does not deny that one can speak in general terms of some neurological processes that are fast, automatic, and often unreliable while others are slower, deliberate, effortful and potentially more precise. (*Thinking, Fast and Slow* (2011) by behavioral economist Daniel Kahneman refers to these categories as System 1 and System 2.) If we see a twisted stick on the forest floor, mechanisms we can imagine within System 1 may make a quick-and-dirty assessment that it is a frightening snake, setting off bodily reactions of fear. With a bit more time, mechanisms we can imagine within System 2 can look beyond the quick shape to the other details of the object and conclude that it is a harmless stick, at which point the bodily fear reaction abates.

What is happening in the brain and body in this threat-detection scenario? The amygdala is a portion of the brain that has a starring role in threat detection. Lying in the midbrain as part of the limbic system, it has as a primary purpose to determine when something threatening is occurring (or no longer occurring) and to send out (or call off) the alarm. The amygdala's primary inputs are from the senses, but also from the prefrontal cortex, the brain region involved in decision-making, control, and social interactions. When the amygdala sees something snake-shaped in a location where one might expect to see snakes, it very quickly puts the body on alert to take defensive action as needed.

The "alert" mentioned above chiefly consists of signals to the hypothalamus, a structure deep in the hindbrain. When the hypothalamus receives enough signals from competing portions of the brain (including the amygdala), it activates the sympathetic nervous system. When the sympathetic nervous system is activated, the body is on high alert. We see more stress hormones (including adrenalin and glucocorticoids like cortisol), increased pulse, higher blood pressure, higher sense acuity directed toward the threat in a form of tunnel vision, and so on. We see less of things that the organism can safely postpone until the threat is resolved, like digestion, defecation, urination, sleep, eating, drinking, sex.

For our purposes, this activation of the sympathetic nervous system is equivalent to what is popularly known as the fight-or-flight response. We are on edge, ready to leap. In fact, this response is better described as the freeze-flee-fight response, since becoming immobile is one of the defensive options available. Indeed, for most prey animals, freezing upon perceiving a threat is the first response; fleeing comes next when it becomes apparent that freezing will not work; and then fighting is a last resort. Each of these defensive strategies is increasingly costly in terms of bodily resources and potentially risky in terms of success. Upon seeing something that might be a snake

The fight-or-flight response can be turned off; pursuing the strategies to their ultimate conclusions is not a given. The amygdala has numerous projections to and from the prefrontal cortex. Even while the amygdala is indicating the presence of a presumptive threat, the prefrontal cortex can undertake the slower, more careful assessment of the threat (taking advantage of the heightened sensory awareness and

tunnel vision). In our example, the prefrontal cortex concludes the stick is harmless. Having reasoned its way to this conclusion, the prefrontal cortex can relay an "all clear" signal to the amygdala and then to the hypothalamus. The sympathetic nervous system deactivates.

The process of calming down the amygdala takes effort, in part because when the amygdala is activated, it sends signals (electrical and hormonal) that can inhibit the prefrontal cortex. ("I'm not listening to you right now, I'm dealing with this threat.") This raises the question of how much control we can assert over our excitable amygdalas when we are insulted.

In some settings, the panicky amygdala takes the lead, reaching conclusions before our conscious brain functions can be invoked. Studies by Joseph LeDoux have shown that there is what calls a "high road" and a "low road" to the amygdala. Condition a mouse to become frightened of a bell (by pairing it with an unpleasant stimulus like an electric shock). Then destroy the connections between the auditory cortex – the portion of the brain that interprets sounds from the ear – and the amygdala. When the bell rings, the mouse still becomes frightened, even though the signal has not gone through the type of System 2 processing that the auditory cortex ordinarily provides. The amygdala, therefore, can make assessments of stimuli that reach it through a low road, and not through the high road of our more complex cognition centers.

However, some stimuli only will only be meaningful when they arrive through the high road. These will include threats that are social in nature. The prefrontal cortex is highly tuned to facilitate social interactions. Species of primates that have more complex social interactions have correspondingly larger prefrontal cortices. The amygdala treats some events as worthy of attention that can only be the result of cues from the prefrontal cortex. Social situations that have been shown to lead to amygdala activation include social exclusion or cheating in economic games. Consider social exclusion. Study subjects have participated in a computer game called "cyberball" where players use a game controller to throw a virtual ball among them. If a subject believes that the other players are excluding her by never throwing her the ball, her brain reacts in many of the characteristic ways that occur when processing other dangerous or injurious stimuli.

When an insult delivers the message "I do not respect you," it makes sense to conclude that some of the same circuits involved in exclusion are also at play. The information that upsets the amygdala – "I am being insulted" – comes from the prefrontal cortex, so calming signals from that same region – "calm down" – can also reach the target. To continue the metaphor, System 1 may act more automatically and faster than System 2, but rarely does a functional person's System 2 become unable to exert influence.

[Include discussion of testosterone and stress hormones.]

***

Part II.A has described the basics of brain functions upon being insulted. The following sections consider two important ways in which those brain functions are subject to social influences.

## B. The Culture of Honor

The fighting words doctrine gives the judiciary the job of deciding which insults are sufficiently enraging to be proscribable. But this is an after-the-fact inquiry. Under what circumstances will an insulted person actually consider the insult to be of such gravity? Judicial opinions that generally support the doctrine presume that it is biologically inevitable that virtually everyone will respond to insults with either violence or at least violent urges. In fact, while the reaction to insults may involve biological processes (as described above), whether the insult is considered threatening is largely socially constructed. Critics of the fighting words doctrine have argued that it is anachronistic (redolent of the Old West) and gendered (describing a predominantly male response to insult). These observations can be further fleshed out by reference to the sociology and biology of cultures of honor.

Sociologists have developed the term "culture of honor" for societies marked by fierce protection of a reputation for manly toughness and powerful reaction to insult. The fighting words doctrine presumes a culture of honor. An insult must be avenged, and done so personally and violently if necessary.

Cultures of honor are often found in mountainous or agriculturally sparse regions. Sociologists describe two basic preconditions for a culture of honor. (a) The society lacks a strong state apparatus and has little rule of law. (b) The society holds much of its material wealth in a form that can be easily stolen by others (typically this means herding communities). In such settings, protecting resources is every man for himself. Deterring a potentially devastating raid by cultivating a reputation for fierceness is highly advantageous. And such a reputation can be readily demonstrated by a violent response to verbal insults. Overall, the frontier society of the Old West is a good fit for this model. So too are some impoverished urban areas, where gang members respond violently to insults.

In their book *Cultures of Honor: The Psychology of Violence in the South* (1996), Richard Nesbit and Dov Cohen sought to test whether the American South fostered a culture of honor, at least amongst its white male inhabitants – and concluded that it did based on a number of measures. In a laboratory experiment that is quite revealing for our purposes, a white male study subject would be asked to fill out a form and turn it in at a desk located at the end of a hall. Halfway down the hall, a beefy confederate of the experimenters would emerge from a room, bump into the subject, and call him "asshole" (presumably one of the fighting words) before leaving. The subjects were tested for their circulating levels of cortisol (a stress hormone) and testosterone before and after the insult. The results showed remarkable differences in the physiological response to insults by Southerners and Northerners. Following the insult, subjects raised in the South saw their cortisol increase by 79% and their testosterone by 12%. Subjects raised in the North

rose only 33% for cortisol and 6% for testosterone. Surveys of the subjects also showed that Southerners reported being more bothered by the insult than the Northerners.

These studies show that insults can lead to biological responses – but that culture and upbringing have a huge influence over what is perceived as a meaningful insult demanding a response.

[Add material about sex roles in the culture of honor.]

## C. Retaliatory Violence Rolls Downhill

Let's assume an insult is perceived as an attack on one's honor, activating the amygdala and putting a person into a fight-or-flight response. Will you fight? This depends on a huge variety of variables, including the proximity of the people involved, their relative sizes, their personalities, whether there are witnesses present, whether they think they can get away with it. But despite all the possible confounding variables in any one setting, a general pattern emerges: violence is far more likely to be directed at people lower in a status hierarchy, when someone believes their status is at risk.

There is ample evidence in other species that when an animal in part of a stable social hierarchy becomes stressed, conspecific aggression is exclusively directed at subordinate animals in the hierarchy. Implant electrodes into a monkey's amygdala and stimulate it, putting the monkey in an agitated state. The monkey responds with aggression to those further down the pecking order, and never with aggression up the pecking order. Similar results have been achieved when monkeys are pumped with huge doses of testosterone. Their aggression is aimed downward, not upward. Similar results have been found with mice.

The aggression seen in these studies is primarily *displaced*. When an alpha abuses a beta, the beta then abuses a gamma. For whatever reason, aggression towards subordinates has been shown to have a calming effect on agitated amygdalas. In these studies, the affected monkeys and mice did not really know why they felt agitated; they just felt agitated and felt that aggression towards a subordinate would help.

What happens if the subordinate is actually the source of the agitation? This would be the case of an insult directed up the hierarchy, which is the category where I place insults to the police. The likelihood of aggression in this scenario, it seems to me, is only heightened.

Begin with whether the insult threatens one's status in the hierarchy. If an insult is delivered from the top down, this tends not to alter one's place in the hierarchy. To the contrary, it simply proves it. An insulted delivered up the hierarchy has the potential to cause the insulted person to lose status, and is therefore a threat of more serious loss. Losing privileges and status is felt more keenly than never having them to begin with. Part of this is attributable to endowment effects, where people instinctively value things they have more than things they could have; they fight to resist loss more strenuously than they would fight to achieve gain.

People facing loss of status are also likely to undertake risky behavior (including resort to violence) to avoid losing it. In studies of economic behavior, Kahneman and Tversky observed what has come to be known as a fourfold pattern when it comes to evaluating whether it is worth taking a risk. If the probability of losing something valuable is high, people tend to make risk-seeking decisions. They would be more risk-averse in other settings. The probability seems highest of losing status as a result of an insult in a culture of honor. All of these incentives combine to mean that retaliatory violence will be more likely in response to an insult from below than an insult from above.

Think of the Jim Crow South, where African-Americans were routinely insulted with epithets by higher-status whites. Their typical response was not retaliatory violence, but submission. But if an African-American insulted a white person with an epithet, retaliatory violence was highly probable. (This asymmetry reflects partly whether the insulted person can get away with the retaliatory violence. In that society at that time, white people could get away with it and black people could not. But I believe there is also a psychological component related to positions in hierarchy and the resultant sensitivity to insult as a threat to status.)

## III. Lessons for Law

Assuming I am right about all this, what should be the lessons for the law of insults?

### A. The Retaliatory Violence Theory Should Be Abandoned

Of the possible reasons for proscribing fighting words, the possibility that the insulted person will respond with violence is unsupportable.

#### 1. Law Should Reflect Behavioral Reality

At a basic level, law that purports to reflect behavioral reality but in fact does not should be changed. This is a behavioral realist message.

On this point, one need look no further than *Brown v. Board of Education* (1954) and its reliance on findings that racially segregated schools causes psychic harm to African-American children.

Another example of the law moving closer to psychological reality – although without the benefit of laboratory studies – is *Cohen v. California*, which found that an antiwar protestor wearing a jacket that said "Fuck the Draft" was not fighting words. *Cohen* explained that the emotional force of words can be as important as their meaning; this insight is consistent with modern science's rejection of mind/body (or cognition/emotion) dualism.

## 2. Culture of Honor and the Rule of Law

The fighting words doctrine presumes that we live in a culture of honor. While some subcommunities within the nation may follow that pattern, it is improper for the law to censor speech in order to uphold that subculture's priorities. If, as Oliver Wendell Holmes said in his *Lochner* dissent, the Constitution "does not enact Mr. Herbert Spencer's *Social Statics*," then neither does it enact a culture of honor.

This is more than simply my own normative baggage, and my Northerner's comfort in a world that does not operate within a culture of honor. A precondition for a culture of honor is a society where the rule of law does not operate. A doctrine that values vigilantism and violent self-help in defense of reputation is fundamentally inconsistent with the rule of law. We have laws against retaliatory violence. (Indeed, the provocation defense is treated only as a mitigating factor and not a complete defense, which shows that our criminal law still considers violence in response to insults to be unacceptable.)

It also bears noting that having a law against fighting words is inconsistent with the culture of honor. The ability to seek complete (and even disproportionate) self-help is central to these cultures. If we wanted to write rules for a culture of honor, they would NOT include a mechanism by which an insulted person could seek the aid of a legal apparatus to attain redress. Honor is restored by demonstrating your ability to deal with an insult on your own, not by playing the victim and calling the police or the prosecutor.

However, the logic of self-help in a culture of honor plays out somewhat differently for police. When they arrest a person who has insulted them, they are displaying self-help by engaging in the aggressive act of arrest. Thus we see yet another way in which the fighting words doctrine uniquely privileges police action against those who insult officers.

## 3. The Fighting Words Doctrine Privileges the Powerful

The fighting words doctrine has the effect of privileging the powerful. Because insults directed at them are the ones that will provoke retaliatory violence, the law makes insults to the powerful a graver offense than insults to others. This is not desirable in an open society that prizes the ability of the individual to challenge authority.

## 4. The Fighting Words Doctrine Cannot Cure The Vagueness Problem It Purports To Solve

As described in Part I, the fighting words doctrine of the 1940s arose in significant part to deal with a problem of vagueness. Statutes made it illegal to say "opprobrious" or "derisive" things in public, yet the First Amendment protects the ability to say things that others abhor. To accommodate both impulses, there needs to be a line separating the forbidden insults from the acceptable challenges to authority. The fighting words line was intended to be that workable boundary.

Of course, it has been anything but – and the brain science shows how it never can be. There are simply too many variables at play to predict a violent response, and those patterns that are discrenable indicate that there is are no universal responses to insults. We can instruct juries to use supposedly objective standards of what how a mythical reasonable person would respond to an insult, but there is no such standard.

## B. True Threats: The Dangers of Freezing and Fleeing

Think of the defendant's statement in *Gooding v. Wilson*: "You son of a bitch, I'll cut you all to pieces." This is an epithet coupled with a threat. For whatever reason, the defendant was charged with fighting words rather than with threats. But the discussion in Part II can be used to show how socially threatening words are processed by our brains. If anything, this helps show why the true threat doctrine is a desirable one.

The reasons for proscribing true threats seem to be so obvious that they do not have to be stated. The leading case, *Watts*, considered it obvious that threats were proscribable, so the only live question was whether the defendant's statement was a true threat. Fight-or-flight insights help explain the basis of the true threat doctrine. The activation of the amygdala and the resultant stimulation of the sympathetic nervous system are the same. The differences are (a) the input that caused the fearful reaction (threat rather than insult); and (b) the response (freezing or fleeing, rather than fighting). The freezing or fleeing that occurs in response to a true threat are inimical to further discourse.

The greater intuitive support for the true threat doctrine, as opposed to the fighting words doctrine, can be seen in the pair of cross-burning cases from the Rehnquist Court. In *RAV v. St. Paul* (1993), the city relied on an ill-fitting and not very persuasive fighting words theory: a cross burning will lead to retaliatory violence that the government has an interest in averting. This is inconsistent with our understanding of how top-down insults work, so it was no surprise that a majority of the Supreme Court did not sign off on the fighting words theory (although the idiosyncratic reasoning the majority used to get there may well have been a surprise). Contrast this with *Virginia v. Black* (2000), where the law proscribed cross burnings with intent to intimidate, i.e., threatening cross burnings. This scenario made both psychological and legal sense, and the Court's differential treatment of the different cross burners showed sensitivity to the facts to discern which was a true threat.

## C. Other Bases for Proscribing Insults

If retaliatory violence is not a suitable basis for proscribing insults, might there be others? For this, we can look back to language quickly tossed out in the course of *Chaplinsky*.

### 1. "No Essential Part"

*Chaplinsky* argued that categories of speech may be proscribed if they form "no essential part of any exposition of ideas" and "are of such slight social value as a step to truth that any benefit that may be derived from them is clearly outweighed." *US v. Stevens* (2010) rejected this formula as a basis for identifying new categories of proscribable speech.

For present purposes, it is relevant to note that of all insults, ones directed at government officials – including police – are those with the greatest connection to First Amendment values of self-government. Yet paradoxically, these are precisely the insults that the fighting words doctrine is most apt to target.

### 2. "Inflicts Injury"

In a very brief aside, the opinion in *Chaplinsky* offered, without explanation, the assertion that fighting words are insulting epithets that "by their very utterance inflict injury." This is a very different rationale than the retaliatory violence theory that controlled the entire remainder of the opinion. No majority opinion of the Supreme Court has pursued what it might mean for an insult to inflict injury. It is beyond the scope of this Article to explore this fuller, but it will conclude with some thoughts on possibilities in light of the relevant psychology.

#### a) True Threats

Lower courts have often pointed to true threats as examples of words that by their very utterance inflict injury. The trick then becomes deciding when an insult conveys a threat. Since the true threat doctrine is already fact-specific and requires a consideration of all the circumstances, channeling threatening insults into that rubric would be justified with no alternations to current law.

#### b) Insult as Injury

Some scholars, including some prominent critical race theorists, have argued that racial epithets should be actionable as fighting words. See Mari Matsuda et al., *Words That Wound* (1993). The fit was never perfect, since these scholars also acknowledged that freezing or fleeing, rather than fighting, was the likely response of a racial minority member confronted with an epithet.

Some scholars studying various forms of uncivil online speech are similarly exploring how messages of social exclusion – as seen in the cyberball experiments – can be said to cause real injury. After all, social exclusion triggers ferocious activity in the same parts of the brain that are triggered by other painful stimuli that tort law acknowledges.

It is beyond the scope of this Article to take a position on these topics. Instead, I would note two items suggested by the topics discussed above.

First, eliminating the retaliatory violence concept and the unhelpful phrase "fighting words" from the lexicon will allow these theories to be judged on their own merits. Better not to force round pegs into a square hole. The alternatives seem to be expanding the round hole of true threats so that it encompasses insults that cause similar psychic harms; or create an independent round hole that does not rely on a the true threat doctrine.

Second, any proposal for a new theory of proscribable words will inevitably come with its own definitional problems, just as there have been chronic problems defining the boundaries of obscenity, defamation, incitement, and the rest. The same problems of vagueness and overbreadth that have plagued fighting words statutes would need to be considered.