When the Law Takes Sides:
Autonomously Weighing Reasons for Expression

Angelo Ryu[*]

Law-making requires the legislator to take sides. This encroaches upon individual autonomy. As a result, the law must tread carefully while regulating expression. This includes hate speech. Despite this, the United States stands alone in its expansive tolerance of hate speech. By examining Scanlonian autonomy and Razian authority, this essay argues that America got it right- proscribing hate speech impermissibly erodes autonomy. Laws that prohibit expressive harm, when limited in scope to harm that targets protected classes like race, presume an answer to a normative debate where that debate is the reason the expression is proscribed. As a consequence, the law interferes with the process of internalization by excluding first-order reasons for action from deliberation.

[*] Working paper, March 4, 2018

1. The Nature of Legal Side-Taking

I disapprove of what you say, but will defend to the death your right to say it. So says Evelyn Beatrice Hall when describing Voltaire's view of free speech.[1] This view remains dominant among liberal theorists. But arguments for tolerating hate speech are under attack. Prominent scholars raise new arguments against the classical call for toleration. My paper defends the liberal position by focusing on the special nature of laws. Like those before me, I ground my argument on individual autonomy. I start with the claim that autonomy, or the freedom to choose, is good and desirable for society. The state must be able to regulate the harmful consequences of an action. What the state cannot do is target the expressive component of speech by discriminating on expressive content. This excludes from deliberation important reasons for action. Doing so strips the state of its legitimacy.

Consider what happens when a law is passed. There is an expectation of altered behavior. The prudential calculus shifts. I must now alter my behavior to conform to the new law. But more is going on. Many times, when the legislature adopts a law, it does so for the purpose of taking a side. By adopting a national religion, for example, the state expresses a message of endorsement. Not because of what the law publically means to the average citizen. Rather, take the perspective of the state. Assume that there exists a norm of plurality. Assume further that there are no countervailing obligations against that norm. The norm of plurality comes with an expectation of equality, of non-preference. Therefore, when the legislature enacts a law that explicitly moves away from that norm, it adopts as its own that particular religion. The state categorizes different religions when it moves away from neutrality. Religion A is "in" while Religion B is "out." If one allows that legislating is a rational activity, it is safe to assume that the legislature did not just pick a random religion off the shelf. The state threw its weight behind the endorsed religion.

Consider instead a proscription of expression. It is different from a declaration of preference because proscriptions of expression negatively rule out acts that could otherwise be taken by their subjects. It is further different from a proscription on non-expressive acts because it targets not just the action, but also the meaning behind the act. Consider how states regulate non-expressive acts. A law against murder punishes the *action* of murder. The varying degrees of murder punish premeditation. In this way, the planning of the crime constitutes a separable act, punishable on its own (conspiracy) or as an aggravating factor (murder in the first-degree). Murder is murder regardless of its motivation. The state can similarly regulate expressive acts. For example, a blanket ban on expressive harm solely targets the act. Consider how a law against defamation punishes the act of expressing a falsehood that causes reputational injury to another person.

Hate speech is different. A law against hate speech targets the meaning of expression. And it does so in a special way. Hate speech legislation targets the thought behind the act as morally blameworthy, not just the harm that comes from expression. The law takes a side- declaring as

[1] S. Tallentyre, *The Friends of Voltaire* (London: Smith Elder & Co, 1906) 198

wrong the claim that people are worth less because of their racial classification.[2] A judgement is made against making certain impermissible judgements. To do so, the law must first adopt the prior premise: A person is not worth less *because* of their protected classification.

This seems uncontroversial. The state must treat its citizens equally. There is, however, a difference between equality from the point-of-view of the state and equality among persons as enforced by the state. The first case applies to state-sanctioned acts while the second involves the state regulating non-state actors. But people judge others unequal all the time. We might judge someone as deserving less respect if they have a lower educational background. We might look down upon people with less money. Maybe factors like immutability play a role, especially when race is involved. We might judge a person with a disability as deserving more generosity and kindness. And if immutability is a controlling factor, does the advent of sex reassignment therapy reduce the desert of sex as a protected class?

To be clear, I am not making a normative argument as it relates to our personal ethics. I am saying that all forms of discrimination are morally equivalent. Rather, I am addressing the issue of whether the state can make that determination for us, choosing which classifications are protected and which are not. By taking a side, I argue that the state must presume an answer to questions of value and worth among individuals. In this paper, I present a defense of hate speech grounded in the classical liberal tradition.

Throughout this paper, I will examine the relationship of harm caused by expressive acts. I find three structural components to such a relationship: the harm-bearer, the harm-agent, and the harm-content. The harm-bearer is the locus of harm, in that it is the particular identity of the person being harmed. The harm-agent is the person from whom the harm originates, having control over the publication of the harmful expressive content. The harm-content is the particular nature of the expression that causes harm.

By regulating hate speech, the state attempts to prevent this relationship of harm. In doing so, the state must first answer questions on the nature of offense. What counts as offense, and what type of offense should be prevented? In the following section, I examine the case of *R.A.V. v. City of St. Paul* as an example of how a principle of neutrality could be applied in a jurisprudential sense against state efforts to answer the questions presented above. [3]

2. Constraining Legislative Authority

26 years ago, the U.S. Supreme Court struck down a Minnesota statute intended to punish cross-burning and other expressions of religious or racial animus. Controversy ensued.[4] After all, the

[2] This need not be limited to race; impermissible categorizations could include religion, gender, or others. However, I will primarily refer to race when discussing hate speech.
[3] *R.A.V. v. City of St. Paul*, 505 U.S. 377, 411 (1992)
[4] *R.A.V.* has sustained vigorous criticism in the legal scholarship. For a semantic criticism, see Mark Facchini and Peter A. Grossman, "Metaphor and Metonymy: An Analysis of R.A.V. v. City of St. Paul Minnesota" (1999) 12 *International Journal for the Semiotics of Law* 215-221.

American courts have long-recognized the conditional nature of the right to free speech. In *Chaplinsky v. N.H.*, the Supreme Court carved out a 'fighting words' exception to the First Amendment, defined as words "which by their very utterance, inflict injury or tend to incite an immediate breach of the peace."[5] Enter the St. Paul Bias-Motivated Crime Ordinance:

> Whoever places on public or private property a symbol, object, appellation, characterization or graffiti, including, but not limited to, a burning cross or Nazi swastika, which one knows or has reasonable grounds to know arouses anger, alarm or resentment in others on the basis of race, color, creed, religion or gender commits disorderly conduct and shall be guilty of a misdemeanor.[6]

In dismissing the overbreadth claims, the lower courts interpreted the statute as solely encompassing fighting words. In *R.A.V.*, however, the majority went a different direction. While Justice Scalia acknowledged that the government need not allow expression falling within First Amendment exceptions, he held that an otherwise-valid law regulating expression wholly within an exception to free speech cannot selectively proscribe aspects of that expression for reasons irreconcilable with the justificatory grounds of the underlying exception- an underbreadth[7] doctrine of sorts.[8] In other words, acts can be expressive for multiple reasons; e.g., saying "I'm going to kill our war-mongering President" both furthers a political message and conveys a threat. First Amendment exceptions, applying to one element of expression, cannot be used to regulate acts based on another element if that basis cannot be reconciled with the original exception.

For example, the government can proscribe threats and limit the proscription's applicability solely to threats to the President, because the justificatory grounds of the exception[9] reconcile with the justification of the constraint on applicability.[10] However, *R.A.V.* held that a prohibition of threats, limited to stigmatizations on the basis of race, ethnicity, religion, sex, sexual orientation, or national origin, fail because regardless of how "vicious or severe" the expression, it must fall within "specified disfavored topics" for the proscription to apply.[11] Thus, the law limits its scope to expression that targets particular classifications. This, the Court held, cannot stand, in part because the statute discriminates in determining which categorizations are impermissible and which are allowed. The Minnesota statute does not successfully reconcile its

---

For a criticism on its constitutional analysis, see Andrea L. Crowley, "R.A.V. v. City of St. Paul: How the Supreme Court Missed the Writing on the Wall" (1993) 4 *Boston College Law Review* 771-801

[5] *Chaplinsky v. New Hampshire*, 315 U.S. 568, 572 (1942).

[6] *R.A.V.* 380

[7] Compare this to the overbreadth doctrine, where a rule would be unconstitutionally overbroad if, in attempting to restrict unprotected expression, it creates a veil of proscription that covers protected expression. *Board of Trustees of N.Y. v. Fox*, 492 U.S. 469, 483

[8] R.A.V. 384

[9] Here, the justification for the law is the physical danger tied to credible threats.

[10] The justification for the limitation on applicability is that the President is especially exposed to physical danger, and that danger to the President is especially harmful to the country.

[11] *R.A.V.* 391

reliance on the fighting words exception with its limitation to expression targeting certain classifications as fighting words are harmful regardless of whether they are expressed to stigmatize based on race, which is protected by the statute, or sexual orientation, which is not protected.

On its own, however, this argument underdetermines arbitrariness. While it shows that the statute fails to protect some who may be at a similar risk of harm to those belonging to protected classifications, this does not seem fatal. It seems good to protect those falling within at-risk classifications, even if it fails to protect others. Again, I turn to the President example. Suppose now the CIA Director lives with an equal danger of physical harm as the President. Suppose further that the statute continues to limit itself to threats made to the President and not the CIA Director, despite the Director living with an equal fear of harm. Even so, a law protecting only the President from threats seems separate from our understanding of discrimination.

There are, however, relevant differences in a law that limits its scope based on race and a law that limits its scope to solely the President. Prohibiting threats made to the President discriminates on content.[12] However, the restriction on applicability is justified solely on the content-independent grounds.[13] This is so even if others, like the CIA Director, have a similar claim to that justification. Therefore, *R.A.V.* allows for some content-discrimination, as long as the limitation on applicability is justified independent of the expressive content. This would ensure that the "basis of the content discrimination consists entirely of the very reason the entire class of speech at issue is proscribable."[14] It may be helpful to think of *R.A.V.* as establishing a three-tiered test.

(1) Does the law target expression because of its content?

(2) Is the proscription selectively applied?

(3) Is that selectivity justified on grounds dependent on normative considerations of worth and value, as opposed to a definitional or intrinsic quality?

The Minnesota statute fails this test. To see why, examine the nexus between the underlying justification of the proscription and the justification behind distinguishing the sub-class from other classifications. The state, in promulgating the Minnesota statute, adopts for itself a message of disapproval towards expression hostile towards specific racial gender, or religious classifications. In doing so, it moves away from regulating expression on the basis of its harmful effect. Hateful expression on its own need not necessarily cause serious harm. Nor is hateful expression the only way such psychological harm can occur. Its underbreadth suggests a regulation of its expressive content on normative grounds.

[12] This is because it examines whether the content of the expression targets the President specifically.
[13] See note 14
[14] *R.A.V.* 388

Is this disapproval just? Many would argue yes.[15] My answer is mixed- yes, but no. It is right to personally disapprove of racism. Even just. I argue, however, that the government cannot maintain such a position and continue to exercise power legitimately. To do so, I use a Scanlonian concept of autonomy to morally justify a limitation on lawmaking authority that prevents a proscription on expression which discriminates on content while also having a content-based justification.

3. Scanlonian Legitimacy

Most everyone agrees that freedom of expression is important. There is, however, an apparent irrationality in holding speech immune from restrictions generally applicable to other acts. In other words, why does an act deserve an exception from the law simply because it is classified as speech? Many arguments for treating speech differently depend on the positive net benefits of protecting either the given expression or the class of acts.[16] Professor Scanlon, however, provides a non-consequentialist reason to distinguish speech from other acts.[17] To do so, he adopts a Kantian framework,[18] defining a legitimate government as "one whose authority citizens can recognize while still regarding themselves as equal, autonomous, rational agents."[19]

To strengthen his argument that autonomy provides a strong reason to not regulate expression, Scanlon posits an extremely weak account of autonomy. To him, autonomy requires only that a person believe himself to be the ultimate authority in deciding what to believe and whether to act.[20] He disregards the exacting claim that one be perfectly rational to be autonomous. Rather, he argues that an autonomous person need only apply his own conceptions of rationality to the decisions he faces, recognizing the need for him to defend his beliefs and actions as consistent with those standards of rationality.

---

[15] C. Sunstein, *Democracy and the Problem of Free Speech* (New York: Free Press, 1995); A. Amar, "The Case of the Missing Amendments: RAV v. City of St. Paul," (1992) 106 *Harvard Law Review* 1039

[16] Many philosophers take this position. E. Barendt, *Freedom of Speech* (Oxford: OUP, 2007) at 15 makes an instrumental argument for personal development. J. Raz, "Free Expression and Personal Identification" (1991) 11 *Oxford Journal of Legal Studies* at 310 argues for expressive freedom on the grounds that it validates lifestyles.
This argument is also made in the legal academy. A. Chen and J. Marceau, "High Value Lies, Ugly Truths, and the First Amendment" (2015) 68 *Vanderbilt Law Review* at 1437-1438 finds that expression instrumental to discovering illegality should be protected. M. Redish, "The Content Distinction in First Amendment Analysis" (1981) 34 *Stanford Law Review* at 113, 119-121, 136 argues that expression should not be regulated because free speech is instrumentally good for democratic governance.

[17] T. Scanlon, "A Theory of Freedom of Expression" (1972) 1 *Philosophy & Public Affairs* 205

[18] While Kantian in spirit, Scanlon's requirements for autonomy are substantially weaker.

[19] Scanlon, "Freedom of Expression," 214

[20] Ibid. 216

This account of autonomy precludes a legitimate state from having subjects that feel obliged to believe the state correct in its decrees, or subjects that are unable to independently deliberate on the merits of an act. From this comes Scanlon's Millian Principle, which he argues is a natural extension to Mill's Harm Principle.[21] It consists of two harms.

> MP1: harms that come from having a false belief as a result of another's act of expression
>
> MP2: harms that come as a consequence of an agent who acts based on a belief, caused by the expression in question, that such an act is worth performing.

According to Scanlon, MP1 cannot justify legal restrictions on expression. Call this the judgement principle. Neither can MP2. Call this the persuasion principle. He argues that restrictions on expression contrary to the judgement principle are impermissible because it prevents subjects from forming false beliefs, meaning the subject cannot rely on their own reasoning, depriving them of their right to independent judgement. He claims that restrictions contrary to the persuasion principle are impermissible because it proscribes advocacy of certain activity, preventing subjects from independently judging whether the advocacy ought to be acted upon.

4. Internalization as an Autonomous Act

I turn towards the concept of internalization to defend the judgement and persuasion principle. I must first note that the use of the term is not without its challenges. Some suggest that internalization be dropped completely from the behavioral science literature.[22] The issues range from failing to define the term to neglecting to measure its effects.[23] The word is used as a catch-all, fuzzily referring to what most consider an intuitive concept. While keeping in mind these reservations, however, I hope to show how internalization clarifies the relationship between expression and autonomy.

Consider Filius, a child born to an average American household. Like other children, his parents sent him off to preschool when he was 4, teary-eyed and worried for how he will adapt to the outside world. While there, Filius plays with the toys on offer, taking a special liking to the building blocks. But when another child came over to help, Filius tells him that "I will do it alone!"[24] A couple of years ago, when his parents helped him up the slide in the playground, Filius gladly accepted their help. But as he develops a sense of autonomy, Filius rejects the

[21] J. Mill, *On Liberty, Utilitarianism, and Other Essays* (Oxford: OUP, 2015) 13
[22] See R. Winch, *Identification and its Familial Determinants: Exposition of Theory and Results of Pilot Studies* (Indianapolis: The Bobbs-Merrill Co., 1962) 28
[23] See a discussion of these issues in E. Campbell, "The Internalization of Moral Norms" 27 *Sociometry* (1964) 391-412
[24] U. Geppert and U. Küster, "The Emergence of 'Wanting to Do It Oneself': A Precursor of Achievement Motivation" (1983) 6 *International Journal of Behavioral Development* 355-369

assistance of others, taking pride in ownership.[25] Fast forward a year. Now, we expect Filius to have matured further. He has begun to accept help for tasks he cannot manage alone, or even tasks that he could have accomplished by himself, like building a house using toy blocks. Sometimes, Filius even asks for assistance, calling upon his teacher to help him.

Our sense of autonomy develops from an early age.[26] As Filius grew up, he started to claim ownership over the things he came across.[27] Eventually, however, we would expect him to accept help for tasks he cannot manage alone. As infancy gave way to childhood, a willingness to cooperate emerged when offered help. This commitment to cooperation is learned, not innate.[28] Through a series of demands and suggestions, Filius absorbed the preferences and expectations of others.[29] Internalization is the synthesis of these preferences into self-endorsed personal standards of behavior. We are inundated with a barrage of social norms daily. Past childhood, these social expectations become marginally less influential, thanks to an already-developed sense of the self. Whether it be conscious or not, most of us have a sense of right and wrong, good or bad, love or hate. But while our internal standards limit the impact of external norms, these norms still have a profound impact on the way we see the world.

Internalization involves the mixing of outside norms with one's rationality through a reasoned deliberative process. This occurs when a socially-emerging norm is freely accepted as a personally-held rule after a reasoned deliberation of its merits. Therefore, internalization can be understood as a morally significant act of autonomy.

I have previously discussed the implications of a state proscription on hate speech in terms of its impact on the harm-agent. It encroaches on the individual autonomy of the harm-agent by preventing the expression of a personally-held belief on the basis of a normative consideration. I now introduce a more controversial claim: A proscription on hate speech harms the autonomy of not just the harm-agent, but also the harm-bearer. Hate speech must be internalized by its target in order for the relevant harm to take effect. When hate speech causes harm without internalization, the harm is physical, not expressive. The law infringes on the autonomy of the harm-bearer by interfering in the harm-bearer's ability to internalize the expression. This seems counter-intuitive. Consider, however, arguments against paternalistic laws. Those, too, are enacted for the purpose of helping its target. The objection comes in the state's interference in the target's autonomy. The relationship that such a law would create between sovereign and subject is such that the state ceases to be a legitimate ruler.

---

[25] For a study on the childhood development of autonomy, see E. Colson and P. Dworkin, "Toddler Development," *Pediatrics in Review* 18 no. 8 (September 1997): 259; to see the similarities between Scanlon's definition of autonomy and how it is generally used in the psychological literature, see F. Power, et. al., *Moral Education: A Handbook* (Santa Barbara: Greenwood Publishing Group, 2007), 35

[26] Ibid

[27] See footnote 31

[28] D. Forman, "Autonomy, Compliance, and Internalization," in *Socioemotional Development in the Toddler Years: Transitions and Transformations*, eds. Celia A. Brownell, Claire B. Kopp (NYC: Guilford Press, 2007) 285

[29] Ibid, 285

To be clear, the harm structure in hate speech is distinct from the harms targeted by paternalistic laws. In the latter case, the harm-bearer and harm-agent are the same.[30] Compare that to laws targeting hate speech, which aim to prevent the harm-agent from publishing expression that stigmatizes a harm-bearer based on certain impermissible classifications. Many of the objections to paternalistic laws stem from the state's interference in the subject's determination that φing is the best course of action for one's own well-being. A determination about the value of φ has been self-adopted after a deliberative process. I argue that a regulation of hate speech interferes with the harm-bearer's internalizing process as the harm-bearer autonomously deliberates on the harm of φ. The state determining whether φ is deserving enough of proscription is an unacceptable intrusion into the harm-bearer's autonomy. By the state declining to make such determinations for other categorizations, its singles out for protection particular categorizations and expresses a lack of respect for the protected target's autonomy. On a daily basis, we make decisions in determining the nature of harm in expressive component. Whether it be insults directed on the basis of family, intelligence, or age, our society is full of expression intended to harm. Racial and religious classifications may be different in terms of weight, history, and significance. What is not different is the autonomous act of determining the harm of φ.

The harm-agent exercises autonomy when expressing hate speech by deliberating on various reasons for action before adopting and publically expressing that view. For the harm-bearer to suffer a content-based harm from that expression, there must first be a deliberation on the content of expression. Hate speech legislation interferes with internalization on both ends, and by doing so infringes upon individual autonomy.

Consider the following two cases of compromised internalization.

> Lying: Say Person 1 expresses A, a lie, to Person 2, such that Person 2 believes A. Person 2 does φ because of a belief in A. Person 2 would otherwise not φ.
> Threatening: Say Person 1 expresses B, a threat, to Person 2. Person 2 is ordinarily averse to φing. Because B shifts that calculus, Person 2 does φ.

Following Kantian lines, Person 1 has committed a moral wrong in both instances by treating Person 2 as a means. As a tool to achieve Person 1's ends. This jeopardizes Person 2's autonomy. The lie prevents Person 2 from engaging in a rational decision-making process, secretly corrupting his deliberative process. Threats, on the other hand, are less like a corruption of the deliberative process and more a unilateral takeover- a coup. In a way, A is worse. Its malice is in the shadows, taking away Person 2's ability to be rational. The lie causes Person 1 to act according to the wishes of someone else. B is different. It does not strip Person 2 of rationality; rather, it strips Person 2 of the ability to make a decision that is truly free. Both interfere with internalization. The lie interferes with internalization because it prevents a reasoned deliberation on the merits of φ. The expressive content of the threat made by Person 1

[30] G. Dworkin, "Defining Paternalism," in *Paternalism: Theory and Practice*, ed. C. Coons and M. Weber (Cambridge: Cambridge University Press, 2013), 25-39

is not adopted as being one's own by Person 2. Therefore, if the threat is the reason Person 2 does φ, Person 2 acts without internalization.

Regulating hate speech similarly infringes upon individual autonomy. By deliberately hiding what the state considers hateful expression, the state has interfered with internalization in a manner similar to both A and B. Regulating hate speech is like a lie because it prevents individuals from knowing how their neighbors value their worth. Lies corrupt internalization by preventing reasoned deliberation. The same goes for a proscription on hate speech. The hypothesized target of the hate speech is prevented from knowing information that would otherwise be expressed. Perhaps she wishes to avoid those who value her worse. Or at least be more vigilant. Had she heard such speech from a neighbor, maybe she would have considered moving to a different neighborhood. Protecting her from hearing such hateful expression does not prevent the other harms that may be inflicted upon her, least of which is the fact that she is unknowingly made to live next to a hateful neighbor. It is similar to threats, because it is the state, not the individual, that makes the decision whether the expression is offensive. Therefore, this decision is not one the individual can fully adopt through internalization, because the decision has already been made by the state. Like a person making a threat, the state takes over the decision-making process by regulating such hateful expression because it must first presume the target will be offended to a degree serious enough to justify such a proscription. It is wrong for the state to regulate hate speech in a way similar to the moral wrong of telling lies or making threats. In the next section, I examine how the particular legal nature of hate speech legislation further interferes with internalization by excluding from consideration reasons for action.

5. Razian Authority

When the state speaks, we listen. Joseph Raz articulates an account of law based on how we reason through moral issues. To him, laws are exclusionary second-order reasons for actions.[31] He distinguishes between first and second order reasons for action. Positive first-order reasons are reasons for doing an act. A belief in helping others provides a positive first-order reason to give to charity. Negative first-order are reasons for not doing an act. A respect for life provides a negative first-order reasons to not commit murder. Second-order reasons are different. They are reasons for, or against, acting on a reason. The Confucian principle of filial piety[32] provides a positive second-order reason to obey the dictates of the father because the son owes him a duty of loyalty.[33] According to Raz, laws constitute negative second-order reasons, which he refers to as exclusionary reasons. Exclusionary reasons are reasons to disregard conflicting reasons for action. If Raz is correct that laws preempt other reasons for actions, laws would constitute an exclusionary reason to not consider competing reasons.

[31] J. Raz, *Practical Reason* (Oxford: OUP, 1999) 35-48

[32] I refer to the respect owed by a son to his father *because* he is his father in Confucian philosophy.

[33] K. Hwang, "Filial Piety and Loyalty: Two Types of Social Identification in Confucianism" (1999) 2 *Asian Journal of Social Psychology* 163-183

Laws, then, do two things. First, laws first provide a first-order reason to act or not act by its demands. But laws also provide a superior second-order reason which *excludes* conflicting first-order reasons for action. When a law demands that one refrain from murder, it exclude from consideration competing first-order reasons for why that person might wish to commit the crime. It is important to distinguish what happens when a reason outweighs competing reasons from when a reason excludes competing reasons. The first is a fair fight, a battle between multiple first-order reasons to see which comes out top. The second is more akin to the home team locking the doors to their stadium, so that the home team can declare victory by default while their opponents sit outside a locked door.

Raz also distinguishes theoretical authority from practical authority by tracking the classical distinction between theoretical and practical reasons.[34] By theoretical reasons I mean reasons for believing something as true. Practical reasons, on the other hand, are reasons for doing or refraining from an act. Someone with theoretical authority gives us reason to believe in what he is saying because he said it; someone with practical authority gives us reason to act according to his commands because he said it. From this comes the understanding that the state, because of its unique position, commands the practical authority necessary to make law, which Raz understood as being exclusionary reasons for action. With this practical authority comes a responsibility to safeguard autonomy. In the next section, I examine why.

6. Expressive Harm and Public Meaning

When the law censures expressive acts, it addresses either its expressive or active component. Is the expression undesirable because of what is said, or how it is said? To find out, I argue that one need only examine the harm-agent and the harm-bearer. Determine if the harm-agent has a cognizable intention to communicate a message. If so, the act is expressive. Determine further if the harm-bearer has a harm that arises from the consideration of the expression. If the harm arises because the harm-bearer internalizes the content of expression, then the act has caused expressive harm. If not, the expressive component of the action caused no harm; the non-expressive component caused physical harm.

If there exists no cognizable intention to communicate a message, I find the act unprotected under a theory of autonomy. When there is a cognizable intention to cause deliberated harm through a communicated message, I find the state limited in its regulatory power. It may proscribe the expression only when the scope of the law is unlimited beyond the justificatory basis for regulating the entire class of expression.

Here I find it necessary to address Anderson and Pildes' distinction between expression and communication. Expression, they claim, is an act that demonstrates a cognitive state of mind, be it an idea, theory, or belief.[35] According to them, this is different from communication, which

[34] J. Raz, *The Morality of Freedom* (Oxford: OUP, 1988) 29; M. Thornton, "Aristotelian Practical Reason" (1982) 91 *Mind* 59

[35] E. Anderson and R. Pildes, "Expressive Theories of Law: A General Restatement" 148 *University of Pennsylvania Law Review* 1506

they argue is an act that intentionally demonstrates a cognitive state of mind to others with the hope that others will understand the act's intended purpose. Like the relationship between a square and a rectangle, Anderson and Pildes claim that while communication is necessarily expression, the opposite is not always so. In fact, they argue that harm done by communication constitute a small fraction of expressive harm.

Consider the case of Adam, a high school student. One day, Adam does not turn in his assigned homework for class. Consider further that his dog really did eat his homework. Adam certainly did not intend to communicate any disrespect or laziness. Nor did he intend to behave in a way consistent with that of a bad student. His teacher, however, takes his act as demonstrating just that. Adam's act was expressive.

Compare that to the case of Kaepernick, the NFL quarterback. Before sitting for, and eventually kneeling to the playing of the Star-Spangled Banner, he formed a reasoned intention to contravene the social expectation that he stand for the playing of the national anthem. He did so for a reason: to express his belief that the U.S. oppresses African-Americans, especially as it relates to police brutality. He intended for his protest to communicate a message to that effect; his purpose in protesting would make little sense otherwise. Kaepernick's act was communicative.

I wish to argue that this distinction collapses when discussing expressive harm done to, and done by, private actors when viewed from the state's point of view. Consider first the harm caused by communication, where the harm done correlates with the intention of the speaker. Here, it seems fairly obvious to think of the harm as being expressive. However, the injury caused by what may outwardly seem like expression is not expressive harm if the harm fails to correlate with the content-value of the act. In other words, if the injury happens simply because the act was done, it is no more expressive harm in the relevant sense than, say, a punch in the face. Content-independent harms are caused without regard to what is said, like the breaking of glass or the annoyance of a loud party. A harm that occurs from an expressive act but does not correlate with its content-value is different. Consider the simple insult "You idiot." Assume that the statement invokes a hostile reaction on the harm-bearer. Presumably, the harm suffered has nothing to do with harm-bearer thinking herself an idiot. In other words, while the harm is connected to the content of expression, the content-harm of the insult is separable from the content-value of the expression. The harm comes from the fact that the insult was said, not from the harm-bearer internalizing the particular content-worth of the insult. This is not expressive harm because the content of expression is not the source of harm. Rather, the locus of harm is the act, independent of its expressive content.

Now, I introduce another conception of harm. One might argue that some expression causes harm because of its public meaning.[36] This could result in the expressive act causing discomfort or a feeling of profound personal injury to unintended targets. This harm exists independent of the original intention behind φ. I believe this argument problematic because it divorces the agent

[36] For this paper, I will use the phrase 'public meaning' and 'social meaning' interchangeably.

from φ in examining harm, while later transferring moral blame onto the actor.[37] Consider again the example of Kaepernick. The intention behind his act, he states publically, is to protest against racially-biased police brutality specifically and black oppression generally.[38] Many, however, take his action to be unpatriotic, insulting to the military, and offensive. Have they suffered from an expressive harm? I suspect many would say no. Perhaps this comes from the intuitive feeling that Kaepernick has not *caused* any expressive harm. It is claimed that harm arising from public meaning does not necessarily rely upon the intention of the actor or the understood meaning of the target. Rather, it comes from what "a competent participant in the society in question would see in that event or expression."[39] Anderson and Pildes provides three instances where the public meaning of φ diverges from its intended meaning.

(1) When the actor is negligent or thoughtless, not considering important reasons for action before acting.

(2) When the actor ignores social or cultural convention.

(3) When the actor unintentionally acts on attitudes or assumptions, like implicit biases.[40]

I deny that φ has a public meaning independent of the speaker's intention or its understood meaning by the intended target. While public meaning is generally used to ascribe an expressive meaning to state action, I address public meaning here because I find it a necessary premise to an argument that the state can regulate an unintended harm done to an unintended target.[41] Otherwise, consider the implications. Without a singular public meaning to ground its understanding, the state is left without a rational benchmark for harm. The state regulates expressive harm by targeting the harm-agent. In the case of Kaepernick, many supported his actions; many were insulted. There was no clear, articulable target for his expression. Must the state take sides, declaring whether it was correct to have been insulted? Or must they make some other calculus, perhaps polling the public? By eschewing the constant of intention, public meaning is the sole anchor that prevents the meaning from diverging into countless directions. Public meaning is necessary to trace the multiple conceptions harm to its source.

I thus examine the viability of public meaning as a means for the state to determine a cognizable expressive harm for the purposes of regulation. Public meaning is a social construction, independent of the intention of the harm-agent or the harm-bearer's understanding. Whether or not φ has a harmful public meaning is determined by the social conditions upon

---

[37] Richard Ekins make a similar argument in "Equal Protection and Social Meaning" (2013) *Legal Research Series*. University of Oxford 5-6

[38] Steve Wyche, "Colin Kaepernick Explains Why He Sat During National Anthem" *NFL.com*, April 27, 2016. Accessed on February 27, 2018 at http://www.nfl.com/news/story/0ap3000000691077/article/colin-kaepernick-explains-why-he-sat-during-national-anthem

[39] C. Eisgruber and L. Sager, *Religious Freedom and the Constitution* (Cambridge: Harvard University Press, 2007) 127

[40] Anderson and Pildes, "Expressive Theories," 1512-1513

[41] Ibid 1520-1528

which the expression resides. I believe that this approach faces insurmountable challenges which prevent the state from legitimately adopting such a standard in determining harm. This is why I limit expressive harm, at least from the point-of-view of the state, to harm caused by a communicated message that correlates with the intention of the harm-agent. To flesh out the problems I see in the concept of public meaning, I examine the three instances Anderson and Pildes provides as moments when public meaning diverges from the intended meaning.

First, there are instances where the actor fails to consider important reasons for action before acting. Consider the case of a trashed hotel room after a long night of partying. Even though the guests did not consider what message their acts would send, it is argued that their inconsiderate behavior expresses disrespect to the neighboring guests (during their loud party) and disdain to the hotel staff in the morning (the trashed room). They have failed to fully consider the reasons for action before acting in such a manner to express a claimed public meaning of disdain.

Next, consider what happens when the actor is ignorant of social or cultural conventions. A classic example would be the unspoken social etiquette codes of different countries. Diplomats must learn to navigate a thicket of complex rules and signals of respect before their assignment. Examples of diplomatic faux paus abound. George H.W. Bush, for example, gave the Australians the equivalent of a middle finger when attempting to flash the V for Victory sign. Michelle Obama touched the Queen of England. It might be claimed that these acts have a disrespectful public meaning.

Finally, public meaning is claimed to diverge from its intended meaning when the actor acts on unconscious attitudes or assumptions. Consider a receptionist working at a hotel that charges a per-occupant room rate. Consider further a businessman taking his female colleague up to his room to continue their discussion. The other day, when he did the same for a male colleague, he was not stopped. This time, however, the receptionist stops them, and demands that he pay for the woman as an additional occupant. The businessman attempts to explain the situation, but the receptionist refused to believe him. The receptionist had acted upon his unconscious attitudes toward women and assumed she was an overnight guest. He has not only embarrassed the woman, but has expressed what might be claimed as an expressed public meaning of debasement, treating her not as a business colleague but a sexual partner.

These examples share a problem of determination. For the state to regulate the first example as having public meaning, it must first decide which reasons for action are important enough to be assumed as the locus of harm. Note that the state need not make such determinations if, for example, it regulates the content-independent harm that arises from a loud noise. In the second example, the state gives legal effect to social and etiquette norms if it regulates the act as having public meaning. When the state assumes those norms as true, it runs into the same problem Minnesota faced in *R.A.V.* Because laws are commands, baking social norms into the many assumptions legislatures make while creating law gives binding force to those very norms. This force is more corrosive in the realm of expression. By proscribing expression that comes as a result of disagreeing or ignoring social norms, the state stops the subject from considering alternative reasons for action. Before the state can find that expression causes public harm because of its social meaning, it must first take for itself large swaths of power. The state must claim itself sovereign in determining not just illegality as commonly conceived, but also in

interpreting norms that are socially enforced. There is recourse if one disagrees with norms enforced socially, and it happens every day. There is a reason we are not all friends. Compatibility with one another often turns on whether their behavior matches our conception of norms, whether it be of kindness, generosity, or attentiveness. We live in a pluralist society. Part of that diversity necessarily means there are differences in thoughts, beliefs, and values. For the state to interpret such norms in a pluralist society, it must take sides and infringe upon our autonomy as exercised in the social arena. The third example presents the opposite problem- that of the state determining the permissibility of personally-held norms. Again, I believe the same problem arises. The state cannot take sides in such conflicts in regulating expression without undermining its legitimacy.

7. Hate Speech and Group Defamation

Waldron argues that hate speech is designed to compromise the dignity of its targets, both in their eyes and among society.[42] This is especially wrong because it targets particularly vulnerable members of society.[43] Therefore, it is a form of group defamation.[44] The rights relationship is structured as such: The harm-bearer has a claim to protection from the state when it comes to group defamation because the state has a duty to ensure that all her citizens are considered deserving of citizenship, and of equal membership in society.

According to Waldron, the individual being targeted is the claim-holder in this rights relationship. This is because Waldron is primarily concerned with individualistic rights, even as he argues that hate speech should be considered group libel. The group is not the harm-bearer. Rather, the individual is harmed by expression suggesting she is worth less because of her group association. However, I find the claim that hate speech is group defamation problematic. First, because defamation law protects everyone, and prohibits lies that cause reputational harm, regardless of how it does so. Second, because defamation law proscribes lies, not normative assertions, valuations of worth, or other forms of opinions. Except in cases where the target of expression is a public figure, defamation law applies evenly to all.[45] It targets the consequence of a defamatory statement- the reputational harm caused by a lie.

Waldron argues that laws proscribing hate speech similarly target only the consequence of the expression. As he understands it, those laws aim to address the reputational harm that comes as a consequence of impugning someone's reputation on the basis of their group association. In other words, hate speech targets especially vulnerable members of society, causing especially serious harm. Therefore, it is not the thought behind the expression that is targeted, but rather its

[42] J. Waldron, *The Harm in Hate Speech* (Cambridge: Harvard University Press, 2012) 5
[43] Ibid 47
[44] While Waldron talks only of libel, I will discuss defamation at large, instead of limiting the discussion to written expression. I believe his arguments apply generally. Also, in the context of hate speech, I am not convinced by his argument that libel causes more serious harm than slander. I argue that a verbal expression of hate speech is no less likely to cause harm than hateful graffiti or an article on a neo-Nazi bulletin.
[45] *New York Times Co. v. Sullivan* (1964) 376 U.S. 254, 279-281

unique consequences. This seems particularly ripe for government regulation. However, laws against hate speech target *how* someone's reputation was impugned on a basis independent of its truthfulness. That it targets the *how*, not the *what*, is important. If this claim of unique consequences is to provide a coherent basis for proscription, it must do more in terms of normative work. To see why, I examine the structural components of how hate speech is regulated.

In the United Kingdom, the Public Order Act proscribes the incitement of all racial hatred.[46] Are there more people who hate any particular race, when compared to another categorization? To see otherwise, simply be a loud Pittsburgh Steelers fan in the company of a drunken crowd in downtown Cincinnati.[47] Of course, however, the difference is *seriousness*. Hatred on the basis of race or religious affiliation is seen as much more serious than sports affiliation.

Stigmatization on the basis of certain classifications seems to have especially worse outcomes. For example, the visibility of race makes it an especially easy categorization to make. And hard to avoid. I argue, however, that a lot of what makes the unique consequences argument convincing is its historical significance. Being racist has historically been tied with especially heinous acts. Even today, many acts are done with racist intent. It seems impossible to not tie the expression to those acts. But the law must do so regardless. Take away for a second the conclusion that acts done with racial intent deserve unique moral blame. Without such a moral determination, the state cannot make a strong enough linkage between the racist expression and the physical act to justify proscribing the expression. People do terrible things all the time, many of which is already proscribed by other content-independent laws.

I return now to this idea of unique moral blame. Compare a law against lynching to a law against murder. Both have the same end effect- lynching is murder, and it is proscribed. A law against murder, however, fails to capture a feeling of uniqueness. It is not intended to censure the specific act of lynching as being especially heinous. I argue that the state cannot target an act as being especially blameworthy based on the act's racial intent. Regardless, by viewing lynching as an act of terrorism, it is clear that there are other avenues for narrowly targeting the type of harm caused by lynching. Lynching causes many to live in fear. This external harm to others, felt by those other than the victim of murder, makes lynching especially harmful. This expressive component constitutes a threat. The *act* of lynching is murder, which is unprotected. I further find the expressed message an unprotected threat. The state is free to target this expressive component on top of a proscription of murder. What the state cannot do, however, is make a judgement that murder is *more* wrong because of its racial intent. There, the state runs into a determination problem, as it must choose which types of expression deserve special moral blameworthiness on its content. The state cannot do so without making impermissible decisions of desert.

Professor Waldron convincingly refutes the argument that the harm of group defamation gets "lost in the numbers."[48] I will make no such argument here. While the size of the targeted class

[46] Later amended to include religious hatred by virtue of the Racial and Religious Act of 2006.
[47] An American football rivalry.
[48] Waldron, *The Harm in Hate Speech* 56

dilutes the harm if the claim is that some members are unworthy, the dilution disappears if the claim is that the association, in and of itself, deserves disdain. By conflating defamation with hate speech, however, Waldron trades on the credibility associated with the confined category of defamation and associates them with hate speech, a category both broader and narrower in scope. This, I think, is a category mistake. Hate speech is more broadly ascribed than defamatory expression because it encompasses claims that are not truly lies- most commonly normative assertions or opinions. It is narrower in scope because hate speech is limited to expression that targets certain protected classes, while defamation can target anyone.

The contours of the category of hate speech rule out many of the arguments supporting defamation. Defamation proscribes only falsehoods. Opinions are not covered. Only lies. Because hate speech applies to expression that in many cases cannot be falsified, the state must make a normative judgement. It cannot do so with a general proposition like "lies are bad" because hate speech cannot be classified as such without a further normative determination. Because hate speech applies only to certain classifications, it lacks the egalitarian quality of defamation law in protecting everyone from reputational harm.

Consider the following statements.

Statement A: "Hitler should have finished the job."

Statement B: "No Blacks Allowed."

Statement C: You people from Mexico are worthless scum.

It is uncontroversial to describe these statements as hate speech. The first is anti-Semitic; the second racist; the third xenophobic. They are, however, not lies. Statement A is an expressed opinion. Statement B carries no independent truth-value. Statement C is a normative claim of desert, which means it cannot be a factual claim in the relevant sense.

Statement A expresses a preference held by the harm-agent. For the state to regulate such expression, it must first directly tread upon the autonomy of the harm-agent to form such thoughts. While the law would not criminalize *having* the thought, the nature of the law would be similar. Consider the case of Adam and Jane, neighbors who have lived next to each other for a number of years without incident. Adam believes Statement A. Jane is Jewish. Take aside for a moment any incidental harm Jane might suffer from Adam's subsequent actions that are tinged by his beliefs. Say that Adam does not harm Jane in any overt way. Say further that Jane does not hold Adam with any special esteem; in fact, despite being neighbors, they barely know each other at all. Now, say that Jane hears Adam expressing Statement A to his friend. Has Jane been harmed? More importantly, does a law proscribing Statement A aim to protect Jane, or target the opinion behind Statement A? I argue the latter. Even if living near a racist constitutes harm, the harm occurs prior to the statement being uttered. And if the law were aimed at protecting Jane, proscribing the very expression that would warn Jane of possible danger seems incoherent. Conceiving of the law as targeting the opinion behind Statement A, I believe, provides a more coherent understanding of what the law targets.

Statement B is a claim. It is, however, not wrong in isolation. The statement carries no independent truth-value separate from the identity of the speaker. Say that Natalie, a government official, announces that "no hats are allowed." One might take that to mean that hats are banned from the country. Now consider Ashley, a billionaire, claims the same. One might expect that to mean that hats are hereafter banned from her extensive private property holdings. Consider further that Bill, a man with no state power or private property, declares the same. One might expect that to mean nothing. The statement itself had no truth-value; the only truth-value that was expressed was contingent upon the identity of the speaker. Therefore, Statement B cannot be a lie. It could very well be incorrect from a legal point-of-view if what it says is in violation of an anti-discrimination statute. The sign, for example, could be placed in an area designated as a public accommodation.[49] In that case, its wrongness is contingent upon a showing of a separate legal command. The law overrules the claim. But it is not a lie per se.

The changing truth-value of expression between Natalie, Ashley, and Bill illustrate why the inner morality of the state fails to fully conform to the ethics of individual action. When Bill takes a side and expresses a position, he provides others a contingent positive reason for action, depending on their evaluation of Bill's credibility. Bill taking a side gives others a reason to also adopt that position dependent on his reputation. When the law takes a side and expresses a position, it provides an exclusionary second-order reason. The law prevents the consideration of first-order reasons. There are two separate distinctions between Bill and the state. First, Bill's expression only provides reasons to people who have chosen to ascribe to him some form of credibility. Second, Bill's expression provides a reason that can be considered, and outweighed, by other competing reasons. The state, however, is in a privileged position when compared to Bill. Therefore, while I argue for the existence of an inner morality of law, I distinguish it from the morality of private persons. This is how I sustain the claim that there are instances where the law is prevented from regulating morally wrong acts when viewed from the point-of-view of the individual, like the expression of hate speech.

Statement C fails to be defamation because it is a normative claim of desert. Claims of desert, however, cannot be lies in the relevant sense. The statement may be morally wrong. But in order for the state to find the statement a lie, it must first take a side. The state must reach the moral determination of falsehood before it can declare C false. Because the state in doing so must violate autonomy, C cannot be a lie in the eyes of the state.

8. Determining Desert

There are already many instances where the state declines to make such a normative claim. For example, while the law generally judges actions, it does not otherwise judge the person. It considers criminal intent, but generally not the reasoning for committing the crime. To see how, one need only turn to a proscription of murder, generally understood as proscribing the intentional killing of another human being.

[49] 42 U.S. Code § 2000

Consider the case of Secel Montgomery Sr.[50] His crime began as a fight over money- his sister-in-law would not give him any to pay for alcohol. With her refusal came a flash of anger, and then a strike. Knocking her out was just the start of something horrendous. Next came the ropes that tied her up, making her helpless. Secel stabbed her to death, on the throat and stomach and chest. He lost track of where his knife plunged, continuing to stab long after her writhing ended. He then washed the blood from his hands. After carefully gathering all the things that could have his fingerprints, Secel called his wife for a ride home. He was later arrested. Afterwards, he was convicted for murder and given a life sentence.

Compare that to the case of Randall Margraves, whose daughters were victimized by Larry Nassar, the former Michigan State University doctor.[51] Randall Margraves was given the opportunity to speak during the sentencing portion of Larry Nassar's trial. He first asked for five minutes alone with the Larry Nassar. The judge refused. He then asked for a minute alone. Again, the judge refused. Losing control, he lunged towards Nassar, intending to take matters into his own hands. It took four officers to stop him. While being led out of the room, handcuffed, he asked what they would do if "this happened to you?"

Imagine what would have happened if Randall was able to carry out his wish in that courtroom. If there were no deputies. No judge to tell him no. All the time in the world to be alone with Nassar. Enough time to kill Nassar. Assume that Nassar was defenseless, and that Randall was in control of his faculties. Randall would have been charged with murder. And he would be guilty. Let us say he is given a life sentence, the same punishment as Secel.

As individuals, we make determinations of desert all the time. Here, it seems easy. Randall does not seem to *deserve* the same punishment as Secel. It is difficult to find Randall morally blameworthy, while Larry Nassar seems to have received his just deserts. Perhaps you believe that you might have done the same thing to Nassar. But the law does not take a side. From the legal point-of-view, both have committed murder. And the state declines to make any such determination of desert, determinations that individuals make freely. I argue that the state should similarly decline to take a side in the case of hate speech. Like in the case of Randall and Nassar, this may lead to uncomfortable results. Results that go contrary to our own sense of desert, or ethics. However, the actions of the state are constrained by factors that do not affect individuals. Autonomy demands that the state decline to make such normative determination. For the state to proscribe Statement C as defamation, it must first decide questions about value and worth. These are questions the state avoids in other areas of the law, for good reason.

Now, I wish to return to the example of lynching. I have until now described its harm as being twofold: the harm done to the victim by the direct act of murder and the harm done to the targets of the act's expressed threat. There is, however, another harm. Consider the aftermath of a lynching. Those who commit the act excuse themselves along racial justifications: The lynching was justified *because* the victim was black. This is plainly invalid as a justification. But this claim is more than simply wrong. It is a claim that the victim was deserving of the lynching

[50] Pam Belluck, "Life, With Dementia," *The New York Times*, February 25, 2012.
[51] Christine Hauser, "Victims' Father Lunges at Larry Nassar in Court" *The New York Times*, February 2, 2018.

because of his race. That the normal rules of social interaction do not apply. Waldron identifies this as a harm suffered by the victims of hate speech.[52] Critical race theorists have found such harm as well.[53] This harm is not unique to lynchings. Note that this is an expressive harm separable from the actual killing of the victim. It results from any act that expresses the view that the harm-bearer is of lower social standing because of a group classification.

9. Changing Obligations

From this harm, Maitra and McGowan derives the conditional claim that, *if* some hate speech has the effect of lowering one's social standing, it should not be protected as free speech. It is important to recognize the limitations of such an argument. They argue that expression having the effect of lowering one's social standing is an "obligation-enacting utterance," which they further argue is not covered under a free speech principle.[54] By obligation-enacting utterance they mean expression that alters one's obligations simply by it being expressed.[55]

They give three examples of such expression.[56]

(1) Donald Trump says, "You're fired" to an employee.

(2) Joan says, "I'll pay you $10 if you remove the snow from my driveway" to her neighbor's son Jimmy.

(3) The boxing instructor says, "Go ahead and punch me in the face" to a trainee.

Consider the statement "You're fired." It changes the legal obligations of its target, as that person is no longer an employee. The employer need not do anything further; saying the statement is enough. The expressive quality serves to further its action component (of firing the employee) and is therefore unprotected under a principle of free speech. The expression constitutes the act of termination.

But Maitra and McGowan do not stop there. They further stipulate that such expression should be unprotected only if it *significantly* alters one's obligations. They must make such a stipulation because they define obligation broadly. So broadly, in fact, that they believe statements like "nasty weather we're having" and "all politicians are liars" enact a change in obligation (to reply) under the "rules of politeness in our society."[57] Under their definition of

[52] Waldron, *The Harm in Hate Speech* 2-3
[53] C. Lawrence III "If He Hollers, Let Him Go: Regulating Racist Speech on Campus" in M. Matsuda et al. *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (Boulder: Westview Press, 1993)
[54] I. Maitra and M. McGowan, "On Racist Hate Speech and the Scope of the Free Speech Principle" 23 *Canadian Journal of Law & Philosophy* 354, 369
[55] Ibid 350-354
[56] Ibid 350
[57] Ibid 352

obligation, the claim that obligation-enacting utterances are unprotected clearly proves too much. So, they limit it to expression that significantly alters one's obligations, whereby significant obligations are ones that apply directly to a matter of law.[58] In other words, if meeting the obligation is either legally required or legally forbidden, that obligation is significant. The obligation is similarly significant if ignoring the obligation is legally required or forbidden. For example, a promise given to participate in a criminal conspiracy entails a promissory obligation that is significant in this sense. Legally-binding contracts also entail a significant obligation.

In that sense, this analysis is similar to *R.A.V.* This distinction is content-dependent in the sense that it is sensitive to what the expression says: "You're fired" and "You're not fired" mean different things, and this change in meaning alters its "obligation-enacting utterance" status because the latter statement would constitute no change in obligation. But this content sensitivity can be easily reconciled with the original grounds for the distinction- that is, whether the statement changes one's obligation. Therefore, I find no problem with Maitra and McGowan's distinction between significant and non-significant obligation-enacting statements.

They argue that, if some hate speech harms one's social standing, such speech should not be protected as a significantly obligation-enacting utterance.[59] Call this their principal claim. I concede that some hate speech harms one's social standing. But I deny such harm constitutes a change in obligation. Notice the work done by 'significantly,' 'harm,' and 'obligation' in their principal claim.

In some instances, their principal claim is obviously true. Recall the cases of Natalie, Ashley, and Bill. Consider further that Ashley is owns a chain of restaurants. Ashley putting up signs saying "Only Whites" in front of her restaurants would go contrary to anti-discrimination law. Because of Ashley's position as owner, her expression carries weight. And like I have previously discussed, this conditional truth-value falls contrary to the law. Therefore, the expression is unprotected. But this is usually not true. Calling someone fat could lower one's social standing. It does not significantly do so because it does not fall contrary to the law. Normative statements are similarly insignificant in the sense given by Maitra and McGowan because its truth-value is a preference. To consider a normative statement significant would require that the law take sides in a province most would agree are outside the law's reach- our thoughts. And, as a matter of positive law, the U.S. has no such thought-crimes.

What if Ashley places the "Only Whites" sign primarily to express her opinion that her customers *should* be white? Here, I argue that such a sign fails to be a normative statement. Text matters. A statement that does not express a normative position cannot be defended on the grounds that it was intended to be. This meaning is not dependent on any particular cultural or social conditions other than the plain language of the statement. It is clear that the sign's intended targets are potential customers. It is also clear that its targets would understand such a statement as being declarative. Because Ashley holds power at her restaurants overruled only by a superior authority (the law), her sign is unprotected.

[58] Ibid 352
[59] Ibid 344

Maitra and McGowan further argue that a statement harming one's social standing constitutes a change in obligation. I find this to be a category mistake. First, because I find that they conflate harming one's social standing with lowering it. Second, because I find that their use of the word 'obligation' too broad to support the claim that obligation-enacting expression should be unprotected. Couching such a claim by limiting it to expression that deals with a 'significant' obligation obscures the nature of obligation.

I argue that one's social standing can be harmed without being lowered. This seems wrong. But in the case of hate speech, this is easy to see. This is because racial classifications, in many instances, is readily visible. Some hate speech harms one's social standing in the sense that it signals to others that its targets are inferior because of their race. But that is all it does. Say that hate speech is directed towards an individual target. It harms the harm-bearer by singling him out. Some (racist) people that do not know him, or ever even noticed him, will have a worse opinion of him. Therefore, his social standing is harmed. But it is not lowered, because a lowering of social standing implies that there is a change in obligation. This I deny. When it comes to the harm-bearer, the racists were never going to follow the dictates of normal social norms. They were neutral *only* in the individual sense of the harm-bearer, but not as to the group. Perhaps one believes this to be a chicken-and-egg problem in that such expression creates racists. But for that to be the case, the expression must be internalized. I imagine the argument would be that the expression causes a racist society, which leads to socially-occurring racial discrimination, which finally leads to a change in social obligations. It seems obvious that more is going on than a simple change in obligation, like in the case of "You're fired." And if the state were to proscribe expression to prevent potential racists from internalizing a persuasive message, it would take a side in a way that would violate autonomy.

Maitra and McGowan further use the word 'obligation' broadly. To them, obligation means anything that obliges. For example, one is socially obliged to respond to a sneeze with "bless you." Or a bank teller is obliged to hand over the money when an armed robber threatens her life. They refer to the former as a social obligation; the latter as a prudential obligation.[60] I disagree with their conflation because I find their previous argument against protecting obligation-enacting statements reliant on a narrower understanding of obligation that excludes both of the above senses of obligation.

Generally speaking, most find two senses of the word 'oblige.'[61] The first sense is present in the claim that we are obliged to swerve the car to avoid a crash with a power pole.[62] The second sense is used when we say that a doctor is obliged to do no harm.[63] If our intention is to not crash, we are obliged to swerve the car in the sense that it is our *only* option. Had our intention been to die, we would indeed not be obliged to swerve the car. Therefore, saying that we are obligated to swerve the car makes no sense, because an obligation entails a normative assertion. No such normative assertion is present here; if you wish to live, you must swerve. The doctor, on the other hand, is surely obligated to do no harm. Regardless of what he wishes, he ought to

[60] Ibid 361, 369
[61] E. Page (1973) "On Being Obliged" 82 *Mind* 283
[62] Ibid 283
[63] Ibid 286

refrain from harming the patient. Therefore, the first sense of oblige does not include a normative contention, while the second sense does.

I argue that the second sense of "oblige," where a statement obliges because of its normative contention, should be further distinguished between separable and inseparable normative contentions. Separable normative contentions are ones where the normative contention is distinct from the issue at hand. An easy example is legal obligation. The normative question of whether one should follow law is separated from the specific act. Moral obligations are harder. Yes, whether an act is immoral depends on what the act is. In this way, it is content-dependent. But this is not the end of the story. It is still separable because moral principles themselves are independent of particular issues. A Kantian would determine moral judgement without regards to an individual issue. An acceptance of Kantian morals would oblige (and obligate) that judgement. And while an act-consequentialist would judge depending on the end result of an individual act, the process itself does not depend on the issue at hand. The moral calculus remains unaltered; the acceptance of such a process would oblige. Compare this to Hedonistic Egoism, which I argue entails no obligations. Rather, it simply obliges us to do what makes us happiest.

Compare hate speech to the statement "You're fired." While hate speech has a normative contention, "You're fired" does not. Further compare hate speech to "I promise to give you $5." Unlike hate speech, this statement changes both legal obligations (as a verbal contract) and moral obligations (promissory obligation). These obligations, however, are separable. Neither legal nor moral principles depend entirely on the issue in question. The change in legal obligation is not dependent on our feelings about the contract. Moral principles are also independent of what we think we should do in a narrow sense. By narrow preferences I mean a preference held without relying on an abstracted principle independent of the specific facts underlying φ. Compare that to inseparable contentions. The normative contention of hate speech resides entirely in its expressive content. Hate speech relies on the normative claim that those of a certain race deserve to have a lower social standing. It is the acceptance of that claim that obliges a racist. It goes no further than that; the claim obliges solely dependent on how we feel about race.

Statements with a separable contention requires no internalization to oblige the listener. Because its normative contention resides in an acceptance of a principle independent of the content of expression, it can alter the obligations of others. Statements with an inseparable contention cannot alter the obligations of others because its internalization requirement attenuates it from the expression. It merely obliges us to do the act. Inseparable contentions, then, lead to a feeling of constraint. It leads to the idea that one is obliged to do φ despite one's narrow preferences. Unlike inseparable contentions, I believe that separable contentions do not constitute harm on its own because it engages with individual autonomy in a way similar to the persuasion principle. Consider the case of hate speech. It can *oblige* someone to change their beliefs about a certain person if their intention is to be racist. But it does not *obligate* them because the normative contention falls wholly within the racist expression. One does not listen to hate speech and believe themselves constrained by an obligation to do a certain action despite one's wishes to not be racist. There is no feeling of constraint like in the case of inseparable contentions. Thus, statements where the attached normative contention is separable cannot alter obligations. Therefore, I deny that hate speech alters obligations.

This line of inquiry has, however, brought to focus an important underlying premise of my argument. I assume that hate speech, when uttered by a general member of society, does not actually lower ones standing. Therefore, my argument does not allow government officials acting in their official capacity to publically express similar sentiments. Nor would this apply to expression that targets a person of an ambiguous or hidden racial identity in a society consisting generally of racists.

10. Expression and Assault

Next consider the objection that hate speech constitutes a harm similar to assault. Scanlon advances such an argument, analogizing expression that causes an unpleasant state of mind to the common law crime of assault.[64] While he worries that these harms are too minor to be recognized by the law, or ephemeral enough to be difficult to establish, Scanlon concludes in principle that "there seems to be no alternative to including them among the possible justifications for restrictions on expression."[65] I reject this understanding. To do so, I argue that Scanlon failed to consider the process of internalization-as-autonomy, which I believe is necessary for psychological harm to be intentionally caused by expression.

First consider Rachel, an African-American woman living in a mostly-white suburb of Boston. She lives a life like that of her neighbors. From taking a morning jog outside in the park to driving up to her local grocery store to stock up on food, her days look indistinguishable from the other members of her neighborhood. This holds true for most days. Some days, however, Rachel suffers indignities unlike those around her because of her race. Like the time when Mark, a man in line behind Rachel at the grocery store, had confronted her. She remembered how Mark accused her of being inferior; a drain on the community. All because Rachel pulled out an EBT card to pay for her food.

Next consider Adam, a veteran returning home after a tour of duty in Afghanistan. While there, he saw acts of unspeakable violence and human suffering. Since returning, Adam suffers from flashbacks of his time as a soldier: the perennial fear of IEDs while out on patrol, the torturous decisions involved with the Taliban's use of child soldiers, and the abrupt ending of some of his closest friends, forged by the shared struggles of war. I argue that this, perhaps contrary to the ordinary use of the word, does not constitute internalization in the relevant sense. By that I mean that the scarring of the mind in the form of PTSD and other forms of violent trauma are not internalized by the harm-bearer because Adam did not, as a result of a reasoned and deliberate process, adopt the violence and trauma experienced at Afghanistan as his own.

Rachel, unlike Adam, suffered an expressive harm. I argue that an unintentional act by the harm-agent cannot have a cognizable social meaning independent of its intended meaning. Adam's harm was unintentional in the sense that the harm-agent did not form a reasoned

[64] Scanlon, "Freedom of Expression," 211
[65] Ibid 211

intention to cause the specific psychological harm done to the harm-bearer.[66] Because it was unintentional, it was not an expressive act. The racially-based insults hurled towards Rachel by Mark, on the other hand, were clearly expression. It is what makes this act so repulsive that I argue mandates its protection from state regulation. I thereby distinguish assault as an automatic impulse from expression as being internalized through a rational process. Adam suffered from assault, while Rachel suffered from expression.

11. Moral Necessity

I have argued in §3 that a necessary condition for the legitimate exercise in the lawmaking power of the state is consistency with the judgement principle (MP1) and the persuasion principle (MP2). Now I argue that the *R.A.V.* decision can be viewed as a series of rules contrary to those principles.

I begin by discussing possible justifications for a proscription on hate speech. First, consider the argument that some expression covered under *R.A.V.* is harmful because it *threatens* an individual's ability to believe he is a full member of society, thus stigmatizing that person on impermissible categorizations. I believe this falls within the judgement principle, as the harm results from a person coming to a belief of stigmatization. For stigmatization to threaten an individual's belief of societal standing, I argue that it must be internalized, whereby internalization is an exercise in autonomy. When compared to direct physical consequences, like waking the sleeping or triggering an avalanche, content-based harms are distinguished as the harm requires internalization and is dependent on what is expressed.[67]

I have thus far argued for expression as a privileged class of actions primarily on the grounds that it has a special relationship with autonomy. Expression is a vehicle for thought, carrying a communicative content. It also has an action component, which can be proscribed. The reason the content component receives special consideration is because it is hard to separate the content of expression from the thought it communicates. The enforcement of thought-crimes seems an obvious violation of individual autonomy. One step removed is the regulation of what thoughts can be expressed. Yet, even here, reasonable proscriptions can be imagined. First, as choosing to express oneself is in itself an autonomous act, a law proscribing the content of expression can still maintain autonomy. One can weigh competing considerations of whether an independently-formed thought should be expressed.

This fails to hold true when a law regulates the content of expression while presuming as true a normative assertion that is itself the justification of the proscription. This presumption prevents consideration of the assertion independent of the law. The content of law fails to provide enough information necessary for internalization. I premise my argument on a conception of law as the

[66] I use the word violent here not in reference to physical violence, but rather in the violence of mental response.
[67] Scanlon, "Freedom of Expression," 410.

reasoned communication of a well-formed legislature intending to change the general body of law.[68]

To isolate the issue, consider the case of Publius, the sole legislator in the fictional state Midlands. When Publius declares chicken noodle soup the official soup of Midlands, his word is law. When Publius opines in his private capacity that chicken and noodles are the two best foods, his words are suggestive to the people of Midlands, but no more. Publius then declares a proscription of expression that claims chicken is inferior because it is an ingredient of the unappealing chicken noodle soup. In doing so, he has mixed his normative judgement of chicken noodle with the descriptive claim that it is the official soup of Midlands. How Publius has done so is relevant. Naming an official soup in Midlands seems to be a state endorsement of the food. In this instance, however, Publius has gone beyond sending an implicit message of endorsement. Rather, by proscribing the expression of a normative judgement on the presumption that the normative judgement is simply false, Publius has, acting in his official capacity, assumed the corresponding opposite opinion to be true. In doing so, he proscribes *conclusions* that might arise from independent consideration. One might, in the course of reasoned consideration, come to believe that chicken noodle is in fact an inferior soup. This, in turn, presents a reason for expressing himself. By preemptively cutting off certain judgements, Publius prevents an independent weighing of considerations. Internalization, as a rational decision-making process, cannot occur properly if the person is not free to form independent conclusions. Thus, a subject cannot allow the state to protect him from internalizing an expression where the basis for that restriction requires an assumed normative assertion.

Why does Publius prefer chicken noodle? The communicative content of the law provides little in normative justification. This is because the normative claim is hidden within the communicative content of the law, being a necessary presumption but not an explicit, justified claim. In some instances, this does not matter. A law proscribing murder need not justify its normative claim, as the act being proscribed does not engage in a normative debate about the wrongness of murder. On the other hand, a law proscribing expression advocating for the rightness of murder must then in its communicative content justify its implied contention. While this can be done when the claim is made explicitly, the normative content cannot simply be made as a presumption when it is regulating expression that falls explicitly within that debate. Without that justification, the subject is required to simply *accept* Publius' normative assertion because Publius has a monopoly on legislative authority. If he was acting in his private capacity, fine. The subject could develop his own reasoned weighing of such a contention with his personal experience. But when it is declared within a law, especially one that targets expressive content, the law itself must justify its normative contentions. A presumption is simply not enough.

The argument I have presented so far has much in common with Professor Dworkin. He argues that a legitimate state cannot proscribe hate speech on the grounds that citizens must have a voice in forming the majority will.[69] That is, the state has no power to impose its will on those who could not express dissent.[70] I view this as another form of the persuasion principle- citizens

[68] R. Ekins, *The Nature of Legislature Intent* (Oxford: OUP, 2012)
[69] R. Dworkin, "A New Map of Censorship" (2006) *Index on Censorship* 131
[70] Ibid 131

of a legitimate state must have the ability to express themselves for the purpose of persuading their fellow citizens. Like Dworkin, I reject instrumental accounts of free expression. By doing so, I avoid the difficulties in establishing consequential causation associated with such accounts. Rather, I find a right to free expression grounded in some basic principle- for me, the principle of internalization-as-autonomy. And here, my differences with Dworkin become clear. Dworkin's account in many instances relies on a conception of group opinion. He describes the worry of a proscription against hate speech as being an intervention occurring "too soon in the process through which collective opinion is formed."[71] Professor Waldron responds to this argument by saying that society is past the stage where a debate over race is necessary.[72] Waldron's argument is compelling if, as Dworkin argues, expressive freedom is justified on the grounds that the majority will must undergo proper deliberation before it is formed if such a will is to be imposed on society at large. Dworkin argues that proscribing hate speech constitutes an impermissible interference with this deliberative process involved with forming a majority opinion, and thus the state cannot do so without foregoing its legitimacy in enforcing laws that go contrary to hate speech; e.g., laws that protect the classifications targeted by hate speech from discrimination and hateful acts. But Waldron is right. If this is the argument that justifies protecting hate speech, surely it is vulnerable in its conception of hate speech proscription as state interference before the formation of a properly-deliberated majority will. I agree with Waldron's objection that the majority will of rejecting racism can be deemed as properly deliberated. It has reached a norm-setting consensus. I read Waldron as saying that, presently, a proscription on hate speech would come *after* the formation of a properly-deliberated majority will against racism.

My account does not suffer from such vulnerability. Unlike Dworkin, I do not rest my argument on how a group will is formed. I focus instead on how individuals, not groups, *internalize* such a group norm. I do not find the imposition of a group will formed without deliberating on opposing views problematic per se. Rather, I find two things problematic: attempting to force the individual to internalize the group norm and presuming as harmful expression that causes harm to an individual only if it is internalized in a certain way. The majority imposes their will on a normative issue by having the law take a side. I maintain that they can do so, but not when it comes to proscribing expression because of the above two issues. By focusing on a basic principle of autonomy that relies solely on a conception of the individual, I avoid Waldron's objection. It does not matter if the majority will is properly formed, if in imposing that group will the law violates individual autonomy by attempting to force objectors to also adopt that will.

Recent legal scholarship has addressed the normative considerations present with regulating expression. Some have argued that courts should consider the harm that come from an internalization of expression when determining whether such expression is protected.[73] I disagree, and hope to have addressed some of the issues in David Han's proposed audience

[71] Ibid 132
[72] J. Waldron, "Dignity and Defamation" (2009) Oliver Wendell Holmes Lecture, Cambridge, MA 1649
[73] D. Han, "The Mechanics of First Amendment Audience Analysis" (2014) 55 *William and Mary Law Review* 1647-1717. (arguing that courts should consider the likelihood of harm coming from an individual processing speech)

analysis. Others have argued that courts should consider lowering scrutiny to expression that does not involve an individual's rational processes. [74] I am sympathetic to that argument. I hope my conception of internalization as a rational process that justifies increased scrutiny addresses some of Rebecca Brown's arguments.

12. Conclusion

The expression of hate presents significant challenges to those it targets. But *R.A.V.* properly proscribes state power to protect individual autonomy. If laws are viewed as being exclusionary reasons for action, laws that regulate expression implicating a normative question by presuming a resolution necessarily interferes with the act of internalization. Because internalization is best understood as an act of autonomy, it must be preserved if citizens are to consider themselves autonomous. The law must allow for expression stigmatizing on the basis of categorizations. To do otherwise would strip the state of its legitimacy.

[74] See R. Brown, "The Harm Principle and Free Speech" (2016) 89 *Southern California Law Review* 953-1010