SHOULD ATROCIOUS SPEECH BE LEGALLY PROTECTED?

Jill Hernandez
Dean, Honors College
Professor of Philosophy and Humanities,
Texas Tech University, USA

jill.hernandez@ttu.edu

Abstract. Some countries limit speech that is likely to incite hate-motivated violence upon a group or breach public peace. Internationally, political tension subsists between free speech advocates and those who want to regulate "hate speech". In countries without prohibitions against hate speech, efforts to limit harm from public speech acts falls to private actors, who feel pressure either to adopt policies to create safe spaces or to allow all speech. This paper refocuses the debate and argues that the current tension between legal regulations of hate speech and cancel culture antagonists misses an entire genre of speech acts that the law should protect its citizens against-- atrocious speech, which yields atrocious harm. The *Atrocity Paradigm*, the non-ideal ethical theory defended by Claudia Card and others, contends that ethics and legal theory should be dedicated to prevent the worst sorts of harms, atrocities. Speech acts which predictably lead to inexcusable, intolerable harm can be distinguished from those which predictably lead to ordinary, or even, hateful wrongdoing. Focusing on atrocious speech allows for legal protections cantered on transmutative harm and inexcusability, and preserves public good obligations to preserve the existence and dignity of oppressed people groups.

Keywords: atrocious speech, hate speech, atrocity Paradigm, dignity, dehumanisation

SHOULD ATROCIOUS SPEECH BE LEGALLY PROTECTED?

Some countries have sought to limit speech that is likely to incite targeted, hate-motivated violence and which can breach the public peace. Internationally, political tension subsists between free speech advocates and those who want to regulate "hate speech". In countries without prohibitions against hate speech, efforts to limit harm from public speech acts fall to private actors, whether persons,

companies, universities, or social media platforms (McLoughlin 2022, 312). Private individuals then feel pressure either to adopt policies to create safe spaces (especially for those who identify with a marginalised group) or to allow all speech so as to protect the free speech enterprise. These passionate opposites are then frequently reduced in the media to bully groups who either 'virtue signal' or scoff at 'cancel culture'.

This paper refocuses the debate, away from definitions and instances of hate speech, and argues that the current tension between regulation of hate speech and cancel culture antagonists misses an entire genre of speech acts that the law should protect its citizens against-- *atrocious speech*, which yields *atrocious harm*. The Atrocity Paradigm, the non-ideal moral theory defended by Claudia Card (2002, 2010) and others[i], contends that ethics and legal theory should be dedicated to preventing the worst sorts of actions: atrocities. Distinct from other, even egregious wrongs, atrocities are intolerable, inexcusable, culpable wrongs that produce systemic, transmutative harm in those who suffer from them. Atrocious harms are not qualitatively worse than ordinary (or even terrible) wrong actions. They are a different genre of wrongdoing altogether, an effect of which is to obviate an agent's ability to experience a great good. Those who theorise about free speech would do well to distinguish between speech acts which predictably lead to inexcusable, intolerable harm and those which predictably lead to ordinary, or even hateful wrongdoing.

Atrocious speech is not a determinate legal category in international law, although Gregory Gordon's (2017) is the first treatment designed to carve out atrocity speech as legally separate from hate speech. He argues that an operationalised legal prohibition against 'atrocity speech' includes four categories: incitement, persecution, instigating, and ordering, and should be implemented through the International Criminal Court. The value of Gordon's work, in part, is that it motivates legal action against a category of speech that is most strongly associated with genocide (United Nations 2025). One challenge for Gordon's particular

articulation of atrocity speech is that there are instances of hate speech which would fall under his categories but would not lead to atrocious harm. Atrocious speech, to avoid the difficulties faced by hate speech legislation, should instead be understood legally in the way Atrocity Paradigm ethicists cast it. Although atrocious speech and hate share the quality that harm results from their instance, atrocious harms are intolerable (they cannot be borne without transmutative harm to the agent) and inexcusable morally culpable wrongs (there is no instance in which they are permissible). The harm that is produced is *atrocious harm*: systemic, transmutative harm that denigrates human dignity and obviates a person's ability to experience a great good.

Focusing on atrocious speech through the Atrocity Paradigm framework, rather than hate speech, allows for legal protections of groups based on a variety of moral factors centred on transmutative harm and inexcusability, and ensures individual liberty for many instances of distasteful, even hateful, speech. Protections against atrocious speech preserve attacks against the existence and dignity of oppressed people groups, while avoiding virtue signalling and cancel culture bullies. The Atrocity Paradigm recognises that atrocious harms are culpable and inexcusable, but it relates both directly to the plight of those who suffer, what private and governmental actors alike should care about.

1. Refocusing on Atrocious Speech

Nature abhors a vacuum, and the same could be said of the law. In the void of legal regulations on speech, private actors and quasi-private agencies are facing escalating pressure to create norms to manage the current social dichotomy between the desire to protect individual free speech rights and a social good interest in facilitating public spaces that are free from the possibility of physical violence. This isn't to say that countries with liberal free speech protections do not regulate speech at all. In the United States, historical

commitments to limiting speech include protecting public morality, restricting labour union speech, limiting the speech of noncitizens, and regulating certain forms of emerging media (Spackman 2021, 42). However, hate speech—an ill-defined concept with multifarious connotations (in fact, in the United States, "hate speech" is not defined in law at all)—can find exceptions in some policies as they relate to fighting words, true threats, and group libel (Gordon 2017, 74). Mostly, hate speech currently resides in the space left by the absence of legal norms.

Unsurprisingly, the extreme implications of what hate speech could connote define the contours of how the public manages hate speech, especially in countries which lack legal policies to do so. On one hand, worries persist that any speech could be deemed 'hate speech'. If beliefs aim at being true and entail commitment, all beliefs have the potential to offend. If any belief could offend, and is pronounced in a manner the listener perceives as maligning or attacking (and maligning or attacking is also perceived as hateful), then all pronounced beliefs risk being perceived as hateful. Maximally, if true, legislating hate speech potentially sets legal guidelines on all speech. Minimally, legislating hate speech sets legal guardrails on any speech except for popular (or in-group, majority-held) speech. On the other hand, proponents of limiting certain kinds of speech point to the inciting influence hate speech can have on agents who hear it—in fact, many argue that a key differentiating characteristic of hate speech is that it does incite violence in people who hear it, "Hate speech is now generally understood as messages intended to incite hatred and/or encourage violence toward a person on the basis of membership in a particular social group" (Hirose et al 2023, 101). There are multifarious historical examples of political hate speech that incited violence. Without regulation against speech that is incendiary, the argument goes, the government seems to formalise and support speech acts that motivate violence.

The United States, infamously, has "promulgated the world's most speech-protective legal regime for repugnant advocacy"

(Gordon 2017, 84), but that freedom has come with dire social and public consequences. In any country like the United States without (or with limited) hate speech prohibitions, the only sanctions that subsist on speech are cultural norms, and cultural norm standards tend to privilege the majority, in-groups. Private actors from minority groups (or groups who are already marginalised) can feel pressure to adopt policies to create public spaces that are free from hate. In contrast, in the interest of preserving individual free speech rights, many in those countries are left to shrug and allow any speech that is neither libellous nor represents a true threat. These passionate opposites are then frequently reduced in the media to bully groups who either 'virtue signal' or scoff at 'cancel culture'. In the United States, especially, the lack of legislation or policy to limit hate speech has resulted in fomentation about "cancel culture". The term first appeared (and became an internet meme) on Twitter in the early 2010s from a group dedicated to issues affecting the African-American community, in which "cancelling" someone connoted a social boycott, a "last-ditch effort designed to hold individuals responsible for hateful speech" (Clark 2020, 89). Proponents of this public boycott technique argue that, in countries in which free speech is a promoted public good, individual agents and private actors must use cancelling as a means to hold people accountable for their speech acts. Absent guiding laws, social justice requires it (Spackman 2021, 9).

One of the strongest proponents of centring legal prohibitions against inciting speech is Jeremy Waldron, who argues against full protection of hate speech based on the erosive impact hate speech has on human dignity. Waldron distinguishes between two harms that are generated from hate speech, "undermining dignity" and "causing offence". Like Joel Feinberg, Waldron argues that even deeply offensive speech typically does not rise to the level of legislative concern. Speech acts which undermine dignity, however, should receive additional legal censure. (Waldron's concept of dignity is "a person's basic entitlement to be regarded as a member of society in good standing") There are reasons to reject Waldron's

view, but for our purposes, the most compelling may be that it falls prey to the problems facing any proponent of hate speech limitation: whether a speech act properly respects a person's basic entitlement to be a member of society in good standing is as subjective as a speech act that causes offensive (even deep offense)—and Waldron rejects offensive harm as a type to be regulated just because it is too subjective.

There may also be good reasons for countries that have regulated hate speech to continue to do so, and perhaps Waldron's "undermining dignity" principle is objective enough to serve as a sound limiting condition on certain types of speech acts. But, the debate can be refocused in a way that preserves our public good obligation to protect certain spaces from violence (and the worst kinds of harm), to preserve the rights of even vile people to express their views, and to legislate to protect minority and oppressed groups. To do so, we first must wrest the conversation away from individual examples of concrete harms (here, of hate speech) and towards a conversation about atrocities. Atrocious harms do not inhabit moral grey zones—they are always wrong and ought always to be prevented. In ethics, the Atrocity Paradigm is a non-ideal moral theory articulated first by Claudia Card, and has been built out to include guidance for how ethics and legal theory can prevent atrocities. By defining atrocities according to the structure from which they emerge (their systematicity) and the harm which marks their sufferers (their transmutativity), scholars and lawmakers can focus on actions which predictably lead to atrocities, and seek to eradicate harms which obviate a victim's ability to create meaning and experience a great good.

Gregory Gordon's excellent efforts to carve out what he calls 'atrocity speech' as a legal basis of limits on the exercise of free speech do not yet engage with the Atrocity Paradigm in ethics. So, prior to engaging with how the Atrocity Paradigm can strengthen Gordon's work, it is valuable to talk about Gordon's unique and significant contribution to the legal philosophy canon. He is addressing legal issues that have pained lawmakers and philosophers

alike since the Holocaust. Gordon directly attempts to provide tools to combat speech that leads to the erasure of people groups and a frayed moral, social fabric. Gordon, a Canadian scholar, effectively draws from his own country's rocky (and often ineffectual) deployment of hate speech legislation to demonstrate a continued (and growing) need for legal clarity and jurisprudence to protect the public interest in safety. He argues that passing legal policies tied to 'atrocity speech' rather than hate speech can help countries that already regulate speech better address the kinds of speech that predictably bring about atrocious harm—and Gordon offers specific types of acts he is interested in prohibiting: genocide, crimes against humanity, and war crimes (2017, 24).

Gordon's legal basis for articulating a framework to prevent atrocious speech begins with the UN's work from 1946-1948, especially the Convention on the Prevention and Punishment of the Crime of Genocide's treaty to establish genocide as a crime that carries individual accountability under international law (2017, 7-9). This UN work was expanded through the 1993 and 1994 Statutes of the International Criminal Tribunal for the former Yugoslavia (ICTY) (at Article 4(3)(c)) and the International Criminal Tribunal for Rwanda (ICTR) (at Article 2(3)(c)). Relevant to Gordon's purposes, the ICTY and ICTR Statutes expanded the Convention's international concerns to crimes against humanity and war crimes. Four legal criteria for prosecutable actions under the Genocide Convention include: incitement[ii], persecution[iii], instigating[iv], and ordering[v]. Gordon supports national and international prohibitions against speech acts which cause genocide, crimes against humanity, and war crimes, and he argues that actions which incite, persecute, instigate, and order these atrocities should be the subject of law. The main difficulty he sees is that the intervening decades since the Genocide Convention have led to a largely fragmented global understanding of what kinds of speech incite, persecute, instigate, and order. (Although Gordon focuses almost entirely on "incitement" in his book-length treatment, the fragmentation problem he dedicates a third of his attention to is applicable to all

four domains of international law on atrocious speech.) Some of the fragmentation problem is a failure of subsequent jurisprudence to normatively develop the ICTR's elemental doctrinal base (as it was intended to do), and some is a result of national courts ignoring the frameworks ICTR and ICTY established for incitement, persecution, instigating, and ordering (Gordon, 2017, 200).

Fragmentation for these frameworks can be generally categorised by the (mainly) epistemic gaps that legal bodies face when applying the ICTR and ICTY guidelines. What does it mean for a speech act to directly incite a crime? (Some courts, for example, have focused on pre-genocidal speech.) Do we have a universally applicable account of what the "public" good is to protect against? Can we determine what it means for speech acts to directly incite genocide? Can contextually dependent aspects of a particular case be considered in legal determinations of guilt? Could we consistently and coherently define and apply a causal clause that sufficiently protects the public? (Gordon, 2017, 186, 207).

It should be noted here that, despite fragmentation and epistemic limitations, it is reasonable to expect some countries to have various motivating reasons to limit certain kinds of speech. Gordon's home country, Canada, has used positive principles in weighing free speech cases, typically by relating expression to three core values: (1) seeking and attaining the truth; (2) participating in democratic institutions; and (3) promoting diversity in forms of individual self-fulfilment (Hutchinson 2023, 687). The efficacy of these principles is limited because these values can conflict, and other values can emerge from social discourse and emerging legal cases. The law, after all, is a living, breathing thing. *R. v. Keegstra* (1990), for example, was a historic Canadian case which upheld reasonable limits on free speech when the willful promotion of hatred would erode the social fabric and threaten shared values. In *Keegstra,* a high school teacher was charged under the Canadian Criminal Code for willfully promoting violence by communicating anti-Semitic statements to his students. The teacher's conviction was upheld by a majority of the Canadian Supreme Court, which ruled that, "The harm caused

by this message run directly counter to the values central to a free and democratic society, and in restricting the promotion of hatred Parliament is therefore seeking to bolster the notion of mutual respect necessary in a nation which venerates the equality of all persons" (Kuhn 2019, 130). The 1990 *Keegstra* Court seemed to presage the cancel culture debate messaging when they urged that jurisprudential limits on some speech were necessary to serve the public good, even when coupled with non-jurisprudential (public) censure.

> Finally, while other non-criminal modes of combating hate propaganda exist, it is eminently reasonable to utilise more than one type of legislative tool in working to prevent the spread of racist expression and its resultant harm. To send out a strong message of condemnation, both reinforcing the values underlying s.319(2) and deterring the few individuals who would harm target group members and the larger community by communicating hate propaganda, will occasionally require use of the criminal law.

Fragmentation and epistemic questions are overcomeable hurdles to an international approach to legislating speech that can lead to three types of atrocities: genocide, crimes against humanity, and war crimes, according to Gordon. The goal of any well-conceived and well-calibrated law with such enormous social impact as that relating to limits on speech should be to reconcile free expression, mass violence prevention, and doctrinal coherence (2017, 24), and Gordon believes that his atrocity speech framework allows him to do so. An issue with Gordon's methodology, however, is that he does not define an atrocity, yet believes his three categories are self-evidently atrocious. In doing so, he treats atrocity like individual concrete harms—but treating atrocious speech as we would individual wrongs makes a category mistake that threatens his framework from suffering the same fate as fragmented hate speech policies. "Atrocious harms" is a separate class of secular evil.

An additional hurdle for Gordon's particular articulation of atrocity speech is that there are instances of hate speech which would fall under his categories but would not lead to atrocious harm.

Atrocious speech, to avoid the tangles of hate speech legislation, should instead be understood legally in the way Atrocity Paradigm ethicists cast it. Although atrocious speech and hate speech share the quality that harm results from their instance, atrocious harms are intolerable and inexcusable morally culpable wrongs, and the harm that is produced is *atrocious harm*: systemic, transmutative harm that denigrates human dignity and obviates a person's ability to experience a great good.

Rather than address the problem of individual concrete harms, the Atrocity Paradigm treats atrocious evils as a class—intolerable, immoral harms that stem from systems of oppression. "Atrocious harms" refers to the category of evils that are culpable, preventable, create intolerable harm, and threaten the great good of someone's life. (Card 2002, 9, 12-13). Card contends these harms typically stem from systems or institutions of domestic, religious, political, and social power. (She has in mind, for example, genocidal rape and dismemberment, psycho-physical torture whose ultimate goal is the disintegration of personality, child pornography, parental incest, slow death by starvation, the explosion of nuclear bombs over populated areas, etc.). Card's list contrasts a bit with that of the International Law Commission, which considers the following to be categories of crimes which constitute either severe human rights violations or inhumane acts (Murphy 2015, 270[vi]). Atrocious harms are a narrower category (even if the intent is the same in drawing up the list) than that given by the ILC.

"Atrocious harms" indicates a genre or class of secular evil that has two main components: its systematicity and its transmutability.[vii] Atrocious harms result from systems of oppression or violence that deprive a person of having access to what is necessary to live a tolerable and decent life, in a way that could never be justified, even by some later good. (The systematicity condition differs from some concrete, individual harms that are on the ILC's list, which need not—and frequently do not-- result from a system of oppression or harm. Murder, for example, is nearly always wrong, but many states reserve capital punishment as a unique form of state-sanctioned

murder for instances in which it is deemed an appropriate punishment. Serious injury to the body causes suffering but might result from legally and morally non-culpable events, such as a natural disaster, or, more commonly, automobile accidents.) Systems or institutions of harm are those which, "in their normal or correct operation will lead to or facilitate intolerably harmful injustices" (Card 2002, 140). Systems of harm aggravate suffering by narrowing possibilities for victims to flourish, and assault human dignity in ways that are specific to the in-groups that are being violated. (Card cites the treatment of Africans during apartheid and racial segregation that resulted in terror, poverty, and degradation as examples, 2002, 103.) What makes a system a *system of harm* is whether it creates the conditions under which there is a predictable, preventable erasure of human dignity through its effects. Although the public often feels powerless to change or bring down powerful structures, Card implores us to evaluate the ways we are personally complicit in facilitating these evil frameworks, and to take responsibility at least for not doing what we can when we can to prevent them and come to the aid of those who suffer from them. Legal regulation (whether at the local or national level) is a fundamental step in changing structures that cause atrocity. Atrocities demand legal recourse; expressions of hate may not.

Whereas the systematicity condition explains the structure from which atrocities occur, transmutativity distinguishes atrocities from lesser harm -- an atrocity, by its nature, transforms people into something wholly distinct from who they were prior to suffering the harm. What makes an atrocity *an atrocity* just is this transmutative property—it erases a person's dignity and divorces a person from what was significantly and uniquely hers. Atrocities "actually disfigure" those who suffer them (2002, 103), at least in the sense that a person's identity (which is built around the ability to interact in social relationships) is altered by the atrocity. "Major historical examples come readily to mind," Card continues, "ghettos and expulsions of the Jews in late medieval Western Europe" (103). Atrocities differ from unjust inequalities (including particular

instances of hate speech), "which would not be evils if they were merely sporadic or isolated incidents in a life otherwise flourishing [that] become evils when they are systematic and come to pervade one's life" (2002, 103). Rather than adjudicate among concrete, specific wrong acts, then, the Atrocity Paradigm contends that the priority for ethics—and the law--should be to eradicate unjust or imbalanced power structures (both locally and globally) that produce atrocious harm or create the conditions under which atrocious harm is produced.

2. Severing Atrocious Speech from Hate Speech for a Jurisprudential Aim

Given the relative newness of the legal and ethical atrocity speech lexicon, it may not be obvious how hate speech acts are distinguishable from atrocious speech, so it takes some unpacking. In hate speech, an undercurrent of fear in the speech act could suggest impending violence, but implicit or explicit threats of violence are not required for an utterance to be hate speech. Rather, an underlying harm of hate speech is that it seeks to 'other' (or delegitimise) the out-group, whereas what makes the consequences of atrocious speech atrocious is its attempt to dehumanise (transmute) the out-group. Consider two different, real-world cases:[viii]

> [A] In post-9/11 New Jersey, an Islamic Mosque is vandalised with a sign that reads, "Jihad Central".
>
> [B] In WWII Germany, a poster depicts a Nazi boot stepping on a cockroach, which is wearing a yellow star of David. The poster (when translated) reads, "Stamp out the infestation."

The two are similar in that they pick out a particular out-group to ostracise—and they do so in a public, shared space in a manner meant to draw attention to the fact that the out-group is being ostracised. The messages are also meant to motivate shared

sentiment based on fear of the out-group. Important differences subsist. [A] describes and [B] prescribes action. The message in [B] is much clearer than in [A]-- so clear that children can understand what [B] connotes. (Indeed, when I visited the Documentation Centre in Munich with my then 10-year-old daughter in 2018, she read the poster and asked if the poster was what Trump meant when he talked about Mexicans in the United States.) [A] suggests people fear the out-group such that reasonable people could be justified in believing [A] could lead to future violence if other contextual features were in place. [B] demands genocidal action, because it includes not only a threat, but an invocation to eradicate.[ix] [A] could be easily deployed in the cancel culture vernacular, whereas [B] couldn't. [A] is an example of hate speech; [B] is an example of atrocious speech.

Hate speech and atrocious speech differ, as well, in the organising principles behind the speech acts. In hate speech, rivalry and antagonism of others take centre stage as a communicative strategy to attack, but in atrocious speech, the leitmotif binding the legal features of the atrocity is a systematic attempt to persecute and dehumanise members of an out-group (Murphy 2018, 1480). For countries that regulate hate speech, the aim of prosecutions tends to be to thwart speech acts which motivate broader group action against a minority population.[x] A challenge posed to hate speech prohibitions comes directly from the cancel culture/virtue signalling bullies: addressing hate speech through the law may promote cancel culture and censorship, while redressing cancel culture may virtue signal and lead to a perception that the government or municipality supports hate speech (McLoughlin 2022, 356). A perverse cycle can ensue. Yet, if we take the Atrocity Paradigm contributions to be relevant, a more proscribed sense of "atrocity" as related to speech acts ensures that neither cancel culture nor advocates against hate speech unnecessarily limit free speech. Atrocities result from systems of oppression, but the systems need not directly be political, nor perpetrated by organised political groups. Any single agent's speech can be identified and limited as atrocious speech if it is

produced from the systematics of atrocity, and the harm produced is transmutative to an individual's ability to create meaning-making.

An example of the systematicity that can produce atrocious speech (and can inform nations as they navigate and legislate between hate and atrocity speech) is organised propaganda. Propaganda itself, of course, is not necessarily either hate speech or atrocity speech. Yet, when propaganda is hate speech, it is common to see the speech devolve from hate speech into atrocious speech as a result of continued and escalating propaganda messaging. Consider the rollout of Facebook/Meta as the only social media platform in Myanmar, following Myanmar's tumultuous shift out of military rule in 2011 (Stecklow 2018). At that time, Facebook was the sole social media platform in the country. Reuters reports that Myanmar's military, the Tatmadaw, posed on the platform as followers of celebrities and other cultural icons to create troll accounts that would be readily followed by the masses (who saw Facebook accounts as an elevated status symbol). The Tatmadaw then used the platform to begin a steady, though successful, propaganda campaign against the country's minority Muslim population, the Rohingya. Facebook, which did not have a single employee within the country, also lacked employees and software that could read Burmese[xi], so the escalating speech went undetected until users outside of the country reported posts and translated the posts into English for Facebook. In a single week, Reuters and the Human Rights Centre at UC Berkeley School of Law collected over 1,000 new posts, videos or comments in the Burmese language that called the Rohingya "non-human kalar dogs" and "maggots" who must be "exterminated" and "fed to pigs". The result of the propaganda was genocide-- 24,000 Rohingya deaths and the largest human exodus in Asia since the Vietnam War, with over 700,000 people fleeing the tiny country in 2017 (Bakali 2021, 54).

To simply say conditions for atrocity include systematicity and transmutativity is not to say that it is simple to identify acts which predictably, in their normal occurrence, lead to atrocity. The Tatmadaw exacted genocide (and were embraced by many in their

country for their atrocities) in part because they strategically took advantage of weak governmental, political, and social structures, while they also capitalised on the rise of a tech platform that had no infrastructure to prevent their speech from becoming atrocious. But utilising the Atrocity Paradigm tools can aid governments and municipalities who want to preserve individual liberties with a tool to identify and curb speech which leads to the worst harms.

There is a jurisprudential aim to the imperative to protect civil liberties and preserve the public good by guiding policy and law with the Atrocity Paradigms' systematicity and transmutativity conditions. It is insufficient to enable prosecutions; the purpose of the Atrocity Paradigm guidance on atrocious speech jurisprudence should be to motivate state action to take measures to prevent atrocious harms from occurring at all. That isn't to say that individual concrete harms are morally or legally insignificant, but the impact of an atrocious harm is that it leaves the sufferer unable to pursue a healthy, meaningful, or dignified existence. If jurisprudence can be directed to redressing the systems which produce transmutative harm and to holding human agents responsible for perpetrating that sort of evil, individuals would be freer to respond to individual concrete harms when they occur.

3. Modifying Gordon's Atrocious Speech with The Atrocity Paradigm

So far, we have been able to see how the cancel culture/virtue signalling bully pulpits have hindered some governments' abilities to legislate and prosecute hate speech in a way that also limits the ability to contravene speech that predictably leads to atrocities. Hate speech, as a broad umbrella concept, encapsulates speech that, while terrible, many would not regulate. Fragmentary and disjointed efforts to create public policy and jurisprudence have led to unsatisfying results even for countries which have enacted laws to limit public expressions of hate speech. Yet, hate speech as a

category is separate from the genre of speech acts which result in the worst sort of harm. Focusing on atrocious speech through the Atrocity Paradigm framework, rather than hate speech, allows for legal protections of minority groups based on a variety of moral factors centred on intolerable harm and inexcusability. It also protects individual speech—even, in many instances, hate speech. Balancing individual speech protections and prohibitions against atrocious speech allows communities to prevent attacks against the existence and dignity of oppressed people groups, while avoiding virtue signalling and cancel culture bullies. The Atrocity Paradigm recognises that atrocious harms are culpable and inexcusable, but it relates both directly to the plight of those who suffer, what private and governmental actors alike should care about.

To date, Gregory Gordon has provided the singular treatment of atrocity speech law. As has been shown, his work is significant in carving out the contours of speech acts that incite atrocities. But his contribution would be improved by utilising the systematicity and transmuativity conditions of the Atrocity Paradigm in ethics. One obstacle for Gordon's particular articulation of atrocity speech is that there are instances of the speech he wants to limit which would fall under his categories of dehumanisation but would not lead to atrocious harms—the same harms he attempts to prevent or limit through his category of "atrocity speech".

Dehumanisation as a basis for hate speech assessments is particularly problematic, for example, in the age of AI and Chatbot-generated content. Whereas numerous studies since Turing have shown that adding a human voice can have an anthropomorphising effect on how humans feel about cars, vacuums, navigational devices, or—more directly—robots, a recent project looked at the dehumanising impact of removing voice from actual humans and replacing it with text (Schroeder and Epley 2016, 1427). The results are fascinating. Absent paralinguistic cues, humans who communicated solely through text were viewed by respondents as "relatively dead or dull, more like a mindless machine than like a mindful human being" (1428). Even if it is true that not all

dehumanising speech is hate speech[xii], imagine the implications of these findings on hate speech jurisprudence. Consider the frequent phenomenon of political advertisements during campaigns, in which still images of out-group members are superimposed with text. In the most heated campaigns, it is common to see ads that superimpose text over a political opponent (or their constituents) to depict them with lower intelligence, moral standing, or (even) citizenship status. Although many would be comfortable labelling such speech "hate speech", those advertisements do not predictably lead to atrocities. Yet, Schroeder and Epley show that such images have a dehumanising impact—similar to hate speech. The out-group pictured is perceived by subjects in the experiment as less than human, or with less desirable human traits, than the in-group. Couple these findings with AI's ability to rapidly produce hate speech content and deep fakes, and speech emerges in which Gordon's categories (i.e., those acts which incite, persecute, instigate, or order) are met without an atrocious speech act being committed. Yet, Gordon's categories only work from a public policy perspective if they differentiate atrocious speech (which should be limited) from hate speech (which should not).

Some might argue that speech akin to that of deep fakes and AI hate speech should be socially limited in non-jurisprudential ways, whether by imploring others to stay off social media or by demanding accountability in limited policy ways, such as holding social media and tech companies financially liable for bot-generated or promulgated content. All of that *could be true*, and still misses the point. Some speech has the form and content of speech that Gordon would like to prevent or limit as atrocious speech, but does not predictably lead to atrocities. Rather, by augmenting Gordon with the systematicity and transmutativity conditions of the Atrocity Paradigm, the difficulty is ameliorated, and Gordon's categories are preserved.

A result of subjecting legal atrocity speech to the Atrocity Paradigm in ethics is that most forms of private speech would not meet both the systematicity and transmutativity conditions of an

atrocity. Descriptive hate speech, even in a public forum, probably does not meet the conditions of an atrocity, either. Many would argue that the Atrocity Paradigm's conditions would not limit enough speech because it would leave many instances of hate speech as legally permissible, and most proponents of restricting hate speech would want deeper restrictions on public speech, especially. The purpose of this project, however, is to motivate action against speech that predictably yields the worst sort of harms, and hate speech, as a category, does not produce transmutative harm. Applying the Atrocity Paradigm's conditions for atrocity establishes atrocities as specifically different speech acts from hate speech. Doing so preserves a country's ability to limit speech that has a deleterious impact on human dignity (and gives them a better tool to protect oppressed groups) while sidestepping altogether the distracting and stultifying debate between the cancel culture and virtue signalling bully pulpits. Focusing on atrocious speech through the Atrocity Paradigm framework, rather than hate speech or an incitement-based atrocious speech framework without the Paradigm, allows for legal protections of groups based on a variety of moral factors centred on intolerable harm and inexcusability. The Atrocity Paradigm should be thought of as an ethical tool available to legal minds to eradicate what is culpable and inexcusable, and support efforts to meet the needs of people groups who suffer, a result that individuals, political groups, municipalities, and private actors should want to ensure.

NOTES

[i] See, especially, Miller (2009) and Bar On (2007).

[ii] In the end, when these additional criteria are tacked on, the existing framework for determining whether hate speech constitutes incitement should consist of seven elements: (1) purpose; (2) text; (3) context (bifurcated into internal— related to the speaker— and external— related to facts surrounding the speech); (4) relationship between speaker and subject; (5) channel of communication; (6) temporality; and (7) instrumentality. Moreover, these criteria can be organised within the larger

conceptual categories of "content" (purpose and text), "circumstances" (context and speaker-subject relationship), and "medium" (communications channel, temporality, and instrumentality). In turn, these categories can help us answer the what/ why (content), who/ where (context), and when/ how (medium) questions related to the speech for determining whether it legally qualifies as incitement. (Gordon 2017, 17)

[iii] Article 7 of the Rome Statute defines crimes against humanity as a series of acts, including persecution, when committed as part of a widespread or systematic attack directed against any civilian population, with knowledge of the attack. 42 Article 7(h) specifies that persecution must be against "any identifiable group or collectivity on political, racial, national, ethnic, cultural, religious, gender… or other grounds that are universally recognised as impermissible under international law." Article 7(2)(g) then defines "persecution" as "the intentional and severe deprivation of fundamental rights contrary to international law by reason of the identity of the group or collectivity …." (Gordon 2017, 10)

[iv] It consists of "prompting another to commit an offence." In other words, the prosecution must demonstrate a causal connection between the instigation and the perpetrated offence. This entails proving that the instigation "contributed" to the prompted person's commission of the crime. (Gordon 2017, 11)

[v] That crime requires a superior/subordinate relationship, issuance of a command to commit an international crime, an awareness that the order would likely lead to commission of an international crime, and a causal link between the order and the commission of the crime. (Gordon 2017, 11)

[vi] The full list includes: (a) murder; (b) extermination; (c) enslavement; (d) deportation or forcible transfer of population; (e) imprisonment or other severe deprivation of physical liberty in violation of fundamental rules of international law; (f) torture; (g) rape, sexual slavery, enforced prostitution, forced pregnancy, enforced sterilization, or any other form of sexual violence of comparable gravity; (h) persecution against any identifiable group or collectivity on political, racial, national, ethnic, cultural, religious, gender, or other grounds that are universally recognized as impermissible under international law, in connection with any act referred to in this paragraph; (i) enforced disappearance of persons; (j) the crime of apartheid; (k) other inhumane acts of a similar character intentionally causing great suffering, or serious injury to body or to mental or physical health.

[vii] In physics, transmutation is the phenomenon in which one element changes into another, typically through a cataclysmic or nuclear event.

[viii] [A] is adapted from an example in Waldron (2012). [B] depicts an actual

WWII poster, on display for educational purposes at the NS-Dokumentationszentrum München.

[ix] Maoz and McCauley (2008) demonstrate the distinct, though connected, relationship between threats and dehumanising factors in hate speech.

[x] Prosecution has mostly been of individuals who have publicly incited hostility towards armed groups or other organisations. See, for example, *Zana v. Turkey*, 1997, in which the European Court of Human Rights upheld the Turkish conviction of Mehdi Zana, and that Zana's free speech rights were subordinate to a social need to keep peace with the Kurdish regions of Turkey.)

[xi] By 2015, the company had four total employees who spoke Burmese, and none of them lived in Myanmar, whose population was 7.5 million at the time.

[xii] It might be, especially if timeless scholarship like Susan Opotow's (1990) is right.

REFERENCES

Bakali, Naved. 2021. "Islamophobia in Myanmar: The Rohingya Genocide and the 'War on Terror'." *Race & Class* 62 (4): 53-71.

Bar On, Bat-Ami. 2007. "Terrorism, Evil, and Everyday Depravity." in *Feminist Philosophy and the Problem of Evil*, ed. Robin May Schott, 195-205. Bloomington, IN: University of Indiana and Hypatia, Inc.

Card, Claudia. 2002. *The Atrocity Paradigm: A Theory of Evil*. Oxford: Oxford University Press.

Card, Claudia. 2010. *Confronting Evils: Terrorism, Torture, Genocide*. Cambridge: Cambridge University Press.

Clark, Meredith D. Sept-Dec 2020. "Drag them: A Brief Etymology of So-Called 'Cancel Culture'." *Communication and the Public* 5(3-4): 88-92.

Gordon, Gregory. 2017. *Atrocity Speech Law: Foundation, Fragmentation, Fruition*. Oxford: Oxford University Press.

Hill, Jesse, and James Fanciullo. 2023. "What's Wrong with Virtue Signalling?" *Synthese*.

Hirose, Kentaro, Hae Kim, and Masaru Kohno. 2023. "A Survey Inquiry into Behavioural Foundations of Hate Speech Regulations: Evidence from Japan." *Japanese Journal of Political Science* 24: 101–117.

Hutchinson, Camden. 2023. "Freedom of Expression: Values and Harms." *Alberta Law Review* 60: 687.

Kuhn, Philippe Yves. 2019. "Reforming the Approach to Racial and Religious Hate Speech Under Article 10 of the European Convention on Human Rights." *Human Rights Law Review* 19: 119–147.

Maoz, Ifat and Clark McCauley. 2008. "Threat, Dehumanization, and Support for Retaliatory Aggressive Policies in Asymmetric Conflict." *Journal of Conflict Resolution* 52 (1): 93-116.

McLoughlin, Stephen. Spring 2023. "Eliminating Fear Speech: How Free Speech Can Address the Dual Threats That Cancel Culture and Hate Speech Poste to Individual Liberty." *San Diego Law Review Rev* 60: 373.

Miller, Sarah Clark. 2009. "Atrocity, Harm, and Resistance: A Situated Understanding of Genocidal Rape." In *Evil, Political Violence, and Forgiveness*, eds. Veltman and Norlock, 53-76. New York: Lexington.

Murphy, Sean D. 2015. "First Report on Crimes Against Humanity." Documents of the Sixty-Seventh Session of the United Nations, 219-270. https://legal.un.org/ilc/documentation/english/a_cn4_680.pdf. Accessed 7/24/2025.

Murphy, Sean D. 2018. "The International Law Commission's Proposal for a Convention on

the Prevention and Punishment of Crimes Against Humanity", *Case W. Res. J. Int'l L.,* 249-252.

Opotow, Susan. 1990. "Moral Exclusion and Injustice: An Introduction." *Journal of Social Issues* 46: 1-20.

R v Keegstra ([1990] 3 SCR 697), SCC Case Information: 21118.

Schroeder, J., and N. Epley. 2016. "Mistaking Minds and Machines: How Speech Affects Dehumanization and Anthropomorphism." *Journal of Experimental Psychology: General* 145 (11): 1427-1437.

Spackman, E.A. 2021. *Justice and Open Debate: An Ideographic Analysis of Freedom of Speech*. [Dissertations and Theses, Brigham Young University, 9032. https://scholarsarchive.byu.edu/etd/9032.]

Stecklow, Steve. August 2018. "Why Facebook is Losing the War on Hate Speech in Myanmar." *Reuters*, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/. Accessed 7/24/2025.

The United Nations. 2025. "Why Tackle Hate Speech"? https://www.un.org/en/hate-speech/impact-and-prevention/why-tackle-hate-speech, retrieved 07/24/2025.

Waldron, Jeremy. 2012. *The Harm in Hate Speech*. Cambridge: Harvard University Press.