Self-Censorship: The Chilling Effect and the Heating Effect

Robert Mark Simpson

Political Philosophy 1/2 (2024): 345-80

Abstract. Chilling Effects occur when the risks surrounding a speech restriction inadvertently deter speech that lies outside the restriction's official scope. Contrary to the standard interpretation of this phenomenon I show how speech deterrence for individuals can sometimes, instead of suppressing discourse at the group level, intensify it – with results that are still unwelcome, but crucially unlike a 'chill'. Inadvertent deterrence of speech may, counterintuitively, create a Heating Effect. This proposal gives us a promising explanation of the intensity of public debate on topics for which there is, simultaneously, evidence of people self-censoring, for fear of breaching speech restrictions. It also helps to pinpoint two problems with existing theoretical analyses of the Chilling Effect: (i) in how they construe the relation between individual- and group-level discursive phenomena; and (ii) how they characterize the distinctively wrongful nature of inadvertent speech deterrence.

In debates over free speech we often encounter warnings about Chilling Effects (CEs). Suppose there is a proposed law that would restrict harmful speech. The CE-based worry is that this restriction won't only deter the harmful speech that it's meant to deter, but that it will also leave people unsure of what they're allowed to say, thus causing self-censorship of lawful speech. The threat of penalties, combined with uncertainty about where the line of permissibility is, leads to a suppression of public discourse in the areas to which the restriction applies. Some Liberals see this as a reason to avoid speech restrictions generally. For others, it's a reason to ensure that restrictions are surrounded with robust caveats, so that people can be confident they won't be punished for saying things they have a right to say.

There are two importantly distinct phenomena that are commonly – and for the most part, uncritically – bundled together in existing discussion of CEs. These are:

1. *Individual Deterrence*, i.e. individuals being deterred from speaking; and

2. *Group Suppression*, i.e. public discourse becoming suppressed or stifled.

The term *Chilling Effect* is typically used to refer to situations in which, by hypothesis, both (1) and (2) occur – where a non-trivial number of individuals are deterred from speaking, and where this causes public discourse to become quantitatively suppressed, and/or qualitatively stifled. Quantitatively, fewer discussions occur on topics related to the speech that's liable to restriction. And qualitatively, in those discussions that still occur, people are warier about expressing their views forthrightly. (When I speak of discourse being *suppressed*, or *stifled*, here and following, I mean to be referring to these quantitative and qualitative aspects of Group Suppression, respectively.)

In existing discussion of CEs, the presumption is that speech restrictions are liable to cause Individual Deterrence, and that this leads to Group Suppression in turn. I'll argue that Individual Deterrence doesn't necessarily lead to Group Suppression – that it can sometimes, counterintuitively, result in discourse being intensified, instead of stifled or subdued. How? By altering the temperamental composition of the discursive community.

I call this phenomenon a *Heating Effect* (HE).[1]

To be clear, HEs aren't some kind of positive inverse of CEs – that isn't the claim. An analysis of this phenomenon won't leave us feeling good (or indifferent) about the potential for speech restrictions to cause Individual Deterrence. But it will help us understand the dangers of this in a different, subtler way. Also, I'm not suggesting that Individual Deterrence always generates HEs at the group level. My working assumption is that Individual Deterrence has different group-level discursive effects in different contexts.

As things stand, however, the concept of the CE makes it difficult to interpret these variable effects, because it invites us to view Group Suppression as the automatic upshot of Individual Deterrence. In explaining how Individual Deterrence can instead cause HEs (i.e. group-level intensification), I'm providing finer-grained tools for interpreting the kinds of discursive phenomena that our CE terminology is meant to be denoting. I'm showing how a dubious assumption about these phenomena – about the relationship between what happens at the individual level, and the group level – is baked into our conceptual framework, and then identifying some good reasons to doubt that assumption.

[1] This label that I'm assigning superficially resembles Daniel Hemel and Ariel Porat's (2019) notion of a *Warming Effect* – i.e. an increase in the quantity and quality of speech – which, they argue, can occur when anti-defamation laws deter false speech, leading audiences to raise their credence in speech generally, and encouraging speakers with important messages to be confident that listeners will believe the truth of what they're saying. Hemel and Porat are exploring cases where the impact of restrictions is to cause more people to speak. By contrast, I'm interested in cases where restrictions lead fewer people to speak, but where this affects the discursive community's composition in a way that intensifies discourse.

That is what's happening in Sections II and III of the article. In Section IV I explain how this account of HEs sheds light on debates over restrictions on discriminatory speech. There's lots of anecdotal evidence of such restrictions causing Individual Deterrence of speech on controversial topics, like racial injustice. But given the vigorous public debate on those same topics, it seems implausible that speech restrictions are suppressing or stifling discussion around them. My account offers a promising explanation of this initially puzzling pair of observations – anti-hate speech laws and the like are (sometimes) causing HEs, rather than ordinary CEs.

In Section V, I present a novel account of what makes Individual Deterrence caused by speech restrictions wrongful. The most well worked-out answer to this question is an old one, from Frederick Schauer, that appeals to speech's putative transcendental value. That account is unpersuasive because it's based on a perfectionistic view of speech's signal importance, e.g. in relation to well-being or justice. Instead, I argue that Individual Deterrence is wrongful because it conduces to dysfunctionality in public discourse (in a sense that I'll explain below), and this is, I argue, something that proponents of all reasonable political views – not just libertarian speech-lovers – have reason to oppose.

First, though, Section I sets the stage for this by expanding on my quickfire sketch of CEs, above, and noting several key points from existing scholarly work on them.

I. Background

CEs occur when a restriction deters lawful speech, because people are anxious about being penalized under the restriction, and uncertain about exactly which speech will be penalized. They happen when "in the course of pursuing legitimate purposes, a law incidentally deters protected expression".[2] CE terminology *isn't* referring to intentional deterrence of unlawful speech. If a law deters people from verbally threatening others, for example, it's presumably doing exactly what we want it to do. CEs involve inadvertent speech deterrence.[3]

In its standard jurisprudential usage – which I will be following – CE terminology doesn't refer to cases in which extra-legal social pressure by itself causes self-censorship. Naturally there are cases in which law combines with social pressure to deter lawful speech.[4] And we may well think these two sources of speech-deterrence are problematic for similar reasons. (This was approximately Mill's view.) In any case, I'm not denying

[2] Kendrick 2013, p. 1673.

[3] Townend (2017, p. 73) says that CE terminology refers to inadvertent *and* intentional deterrence of speech. This is true of some informal usage of CE terminology. But in the scholarly jargon, the scope of CEs is ordinarily limited to inadvertent deterrence.

[4] For example, in her inquiry into the effects of 'cancel culture' in academia, Pippa Norris (2023) uses the language of chilling to refer to the suppression of conservative opinion due to a combination of social pressure and institutional speech restrictions.

the reality – or badness – of social-pressure-caused deterrence of speech.[5] It's just that my interest for present purposes is in deterrence caused by formally speech-restrictive laws and policies.

We could limit CEs' definitional scope to only include deterrence caused by state-enforced laws. But it seems apt to also include deterrence caused by non-state policies restricting speech, if and when these consist in formally-specified duties or prohibitions, backed up by penalties, like a university's speech code, or a company's ban on staff publically criticizing it. Although such penalties aren't state-enforced, the risks they create, and the wariness this breeds, seem liable to elicit the same type of deterrence that defines our target phenomenon. Having said that, while my use of CE jargon isn't limited to state-enforced legal restrictions, I will, for ease of expression, speak of laws / lawful speech throughout.

Defamation law lies at the center of scholarly work on this topic. It's harmful to falsely traduce a person's reputation, so presumptively the law must provide a remedy for this, either via a criminal prohibition or tort action.[6] However, as Leslie Kendrick says, outlining the CE quandary, if we make speakers liable for all false defamatory statements, we'll deter true speech, because "people might hesitate to speak unless they are certain about the truth of their statements".[7] Protected speech can thus be deterred by the regulation of unprotected falsehoods. This is exacerbated by uncertainties about law's reliable administration, about 'defamed' people suing truthful speakers, and about the potential costs of defending oneself. Given all this uncertainty and risk, speakers may think that it's best to stay quiet. Truthful speech that hurts people's reputations can therefore be suppressed or stifled, despite it being speech that in principle merits protection under free speech norms.

What should we do about this? The standard answer is: we should design anti-defamation laws in a way that tries to mitigate uncertainty-caused deterrence. One technique is to build in protections for some instances of false defamation. In US law this is done via the *actual malice* rule, under which speakers cannot be penalized for defamatory speech unless they either know it is false, or recklessly disregard its possible falsity.[8] Protections for falsehoods that are stated in good faith, neither negligently or recklessly, help uncertain voices to speak up. "By drawing the… line between protected and unprotected

[5] Although see Section IV below (and note 35), for further discussion around this assumption.

[6] Much of the CEs literature seeks to identify how deterrence is elicited by different kinds of defamation law, and how to mitigate this. Influential scholarship includes: in relation to US law, Shiffrin (1978), Youn (2013); in relation to UK libel law, Barendt et al. (1997); in relation to Australian defamation Law, Kenyon (2006).

[7] Kendrick 2013, p. 1637.

[8] The rule is from *New York Times Co. v. Sullivan* 376 U.S. 254 (1964). UK libel law is sometimes adjudged inferior to US defamation law, from a liberal perspective, given its lack of something equivalent to the actual malice rule, which secures breathing space for 'true' defamation; see Kenyon (2006, pp. 9-20). The UK's *Defamation Act* 2013 sought to remedy this, by establishing new defenses against defamation lawsuits, including 'honest opinion' and 'public interest' defenses, but their efficacy in counteracting CEs remains unclear; see Jones (2019).

speech prophylactically," Kendrick says, "courts create 'breathing space' for expression that is truly protected."[9]

In addition to these defamation-related cases, the potential for CEs exists in many areas of speech-restrictive law and social policy, including laws regulating protest, political dissent, offensive and indecent speech, and the dissemination of dangerous / classified research. As Cass Sunstein says, pretty much any significant penalties for false speech can deter truthful speech, given speakers' uncertainty about their statements' truth.[10]

In all these areas, the risks speakers associate with restrictions can result in Individual Deterrence via multiple, complementary routes. Schauer presents a catalogue of these deterrence mechanisms in his seminal account of CEs. One common source of uncertainty and risk lies in the fact that speech restrictions can be applied erroneously. Being aware of this possibility, someone governed by a rule that forbids saying *p* may be discouraged from saying some p-adjacent thing, *q*, for fear of being mistakenly adjudged to have said *p*.[11] In some instances the misapplication of a rule happens because the arbiter isn't well-equipped for their adjudicatory task, e.g. like the university administrator who plays *de facto* magistrate in applying campus speech codes. But even where rules are administered by well-trained judges, there's still potential for human error – and for a rule's misapplication – thus casting a shadow of anxiety over speakers' communicative choices.

When speakers are found to have infringed a speech restriction, penalties typically follow, and these can be seen as an additional deterrent factor.[12] Anyone may worry about facing a mistaken charge. But if the associated penalty is a heavy fine, then someone who knows they can't afford to pay that fine has an extra fear – beyond the initial fear of being branded a rule-breaker – that further deters their speech. The costs involved in defending oneself against such charges create an additional source of deterrence, leading to "a fear of the entire process, with a commensurate increase in the degree of deterrence".[13] And then other non-financial costs (e.g. stress, time) can have additional deterrent effects too.[14]

To sum up: otherwise justifiable restrictions on speech, aimed at deterring harmful speech, are also liable to deter permissible speech. Even if your intended speech is entirely lawful, relative to restriction R, you can be uncertain about whether your speech

---

[9] Kendrick 2013, p. 1637.

[10] Sunstein 2020, pp. 400-403.

[11] Schauer 1978, pp. 694-695.

[12] Ibid, pp. 696-697.

[13] Ibid, p. 700.

[14] Ibid, p. 697. Naturally, there are positive motivations/incentives too, including ways speech can benefit speakers, which models of deterrence must factor in. Schauer suggests that we can model all these factors via an equation: "deterrence = risk aversion ((probability of punishment × extent of punishment) – expected benefit)". Ibid, pp. 697-698 (and note 62).

will incur penalties linked to R's application. Further inquiry into exactly how restrictions influence people's willingness to speak may influence the measures we use to mitigate this. But the shape of the problem suggests the shape of the solution. Speech restrictions must be formulated precisely, and should err towards underinclusivity – only proscribing speech that's clearly liable to proscription. Speakers need breathing space and buffers.[15] This is the standard account of what CEs are, and what to do about them.

Most scholarly work on CEs looks at speech deterrence specifically. In a recent article Jonathon Penney criticizes the speech-centric approach. He argues that inadvertent deterrence influences non-speech acts as well, that all kinds of factors (not only legal restrictions) cause this, and that these factors cause "not just a deterrence effect, but a shaping effect".[16]

These are fair observations, but they don't obviously support Penney's proposal to radically extend the definitional scope of CEs, to include a wide range of factors influencing a variety of activities. This would transform CEs into a byword for anything that influences norm-compliant action. Indeed Penney seems to recognize this and welcome it. He says a CE is "best understood as an act of compliance with, or conforming to, social norms in [a particular] context".[17] I don't deny that we have reasons to examine all the factors that influence norm-compliant action. But we have a distinct term for CEs because, by hypothesis, we also have reasons to zero in on this specific form of norm compliant action – the one that occurs when speech restrictions deter lawful speech. To give this phenomenon a special label isn't to deny that behavior is influenced by many factors (not only laws), and that all behavior (not just speech) is thus influenced. It's to say that there's a particular form of this generic phenomenon that merits special attention. And Penney's account doesn't give us any real reason to believe otherwise. So, *pace* his critique, I'll be working with a standard definitional scope from this point. CEs consist in the inadvertent deterrence of – rather than any sort of *influence upon* – acts of speech – rather than acts of *any sort*.

II. Feedback Effects

As I said at the outset, the CE-related scholarly literature bundles two things together:

1. *Individual Deterrence*, i.e. individuals being deterred from speaking; and
2. *Group Suppression*, i.e. public discourse becoming suppressed or stifled.

When an author claims that a CE is occurring, they are not only positing a decrease in the quantity of speech that is performed, due to Individual Deterrence. They are also,

[15] The *breathing space* metaphor is from Kendrick (see note 9); the *buffer* metaphor is from Schauer (1978, p. 685).

[16] Penney 2022, p. 1455.

[17] Ibid, p. 1488.

typically, positing a qualitative change in the discursive milieu where those acts occur. They're saying that in the affected community, people's impetus to converse becomes somehow stifled and subdued.[18] This can take various forms. It may simply be that fewer discussions and written dialogues take place. Or it may be that in the conversations that do occur, any deep, earnest delving into perilous topics is avoided. The point is that something transpires in the group's communicative dynamics – something that (putatively) arises out of individual-level risk-avoidance, but isn't identical with it. Public discourse is stifled. This idea of how things go at the group level is reinforced by CE jargon's evocative thermal metaphor. Talk of chilling conveys a sense of how the communicative climate feels in the wake of speech deterrence. It conjures an impression of chats becoming frosty. Fewer people want to speak, but also, where people do speak, things are less free-flowing. The dialogue freezes.[19]

It isn't ridiculous to suppose that Individual Deterrence and Group Suppression would go together as a rule. I think this is a mistake, but it takes some reflection to see why. After all, discourse is produced by groups of individuals. If fewer people are willing or

---

[18] Consider three examples that support this claim about how CE terminology is typically used.

- In their analysis of libel law and the media, Barendt et al. (1997, pp. 189-194) distinguish *direct* CEs, where media actors self-censor for fear of incurring legal penalties, from *structural* CEs, where actors refrain from addressing particular topics in anticipation of the pressure to self-censor. In both cases, the authors argue, libel-law-triggered deterrence subdues discourse *generally*. It "narrows the range of what is thought publishable" and "remove[s] certain topics altogether from exposure" (Ibid, p. 192).
- Thomas Hazlett and David Sosa (1997) discuss CEs in relation to the *Fairness Doctrine* (FD): the US law that applied from 1949-87, which required broadcast license-holders to offer equal access to rival political viewpoints. Hazlett and Sosa examine data which, they say, show that the FD had a deterrent impact on the broadcasting of specific contents and on the adoption of program formats. Their claim isn't just that individual licensees were deterred from making these programming choices. It's that this added up to a *systemic* reduction in the presentation of controversial viewpoints in American broadcasting, while the FD applied.
- Turning from scholarship to case law, in *Citizens United v. Federal Election Commission* (558 U.S. 310, 2010), the core justification for deregulating campaign finance constraints was that these constraints interfered with free exchanges in the marketplace of ideas. The argument wasn't just that particular associations were discouraged from conveying their views under these constraints. It was that political debate was being stifled in a further-reaching sense. The majority's opinion, expounding this justification, mentions CEs 23 times. Similar reasoning appeared in the more recent case *Counterman v. Colorado* (600 U.S. 66, 2023), in which the court's decision – that a *mens rea* of recklessness must be shown, in order to place threatening speech outside of First Amendment protections – was justified via a mixture of (i) plausible observations about how legal restrictions on threats deter individuals from engaging in hostile speech, alongside (ii) more speculative concerns about the broader stifling of public discourse.

My point is that in these analyses – and I could cite others to the same effect – the claim that a CE has occurred doesn't reduce to the claim that individual speakers have self-censored. Talk of 'chilling' *can* be used as a way of adverting to instances of Individual Deterrence. But it usually goes further. Usually, talk of CEs is simultaneously adverting to some Group Suppression that is the (alleged) cumulative consequence of many instances of Individual Deterrence.

[19] The linking together of concerns about Individual Deterrence and concerns about Group Suppression is also subtly evident in Schauer's classic account of CEs, The fundamental problem with CEs, for Schauer, is that "something that 'ought' to be expressed is not"; this is bad, "not only because of… the non-exercise of a constitutional right, but also because of general societal loss which results when the freedoms guaranteed by the first amendment are not exercised" (1978, p. 693). The issue with CEs, for Schauer, isn't just that people don't feel able to exercise their rights, it's that the non-exercise of these rights leads to a *general societal loss*. Schauer is suggesting that the badness of Individual Deterrence can only be properly appreciated when we zoom out from the individual, and see how this deterrence translates into a stifling of discourse at the group level.

able to engage in discussion, then, all else being equal, discussion seems bound to quiet down. Discourse naturally fizzles out when fewer individuals want to talk.

What's missing in this surface-level analysis, though, is an account of how deterrent factors interact with the variable traits of the people involved in communicative exchanges, in ways that have a more complex impact on the quantity and quality of those exchanges.

Here is an initial toy example that illustrates one feedback effect of this kind.

> BLOWHARD. Suppose there are three people, Bill, Cara, and Dev. Bill is a blowhard. When he speaks more, this discourages Cara and Dev from talking, because Bill talks over them, which they find irritating. Suppose these people receive an incentive to speak more in some communicative setting, e.g. say they're taking a university class, and they're told their grade could be increased if they participate enthusiastically in class discussion. If Bill reacts positively to this incentive, this may deter Cara and Dev from speaking, as the irritation at Bill's volubility outweighs the incentive of improving their grade. Bill speaks more, but the others speak less, and so the total quantity of discussion – whether counted in terms of token utterances, or interlocutory interactions – decreases.

Here's another toy case, loosely resembling the Heating phenomenon sketched above.

> PEACEMAKER. Suppose Pat is a peacemaker – someone who encourages others to get along, and to respond constructively to each other's speech. When a dispute is burgeoning, Pat's instinctive reaction is conciliatory. She wants to calm things down so that the conversation doesn't boil over into conflict. Suppose also that Pat strongly prefers to adhere to civility norms. Her peacemaking traits are, let's say, a symptom of a broader conflict-aversion – a trait that makes her want to abide by the rules of civil dialogue.

Given this combination of traits, Pat's participation in a discussion may reduce the quantity of communicative interaction, by subduing a certain type of verbose conflict.

> PEACEMAKER (CONTINUED). Pat, Quinn, and Rex are enrolled in a university class. They are told that their grade may be reduced if they're uncivil to others in class discussion. Pat hates the idea of breaching a civility norm, and with the new rule in places she feels anxious about how her purportedly civil contributions to discussion could be misconstrued by her peers and professors. So she loses the nerve to say anything besides banal pleasantries. But Quinn and Rex don't share that hang-up, and aren't put off speaking by the rule. With Pat withdrawing herself from the substance of any discussion, Quinn and Rex get sucked into the sort of verbose conflict which, had she not withdrawn, Pat would be helping to calm down. Pat talks less, but this is outweighed by Quinn and Rex talking more.

In both examples, the way that the incentive-shifting rule affects the quantity of discussion is – due to the various speakers' combinations of traits – different to what we might naively anticipate. At an individual level, the rule in BLOWHARD incentivizes more speaking. However, feedback effects, primarily driven by Bill's blowhard personality, result in a decrease in the quantity of discussion. Conversely, at an individual level, the intervention in PEACEMAKER discourages expression. People are given an incentive to

refrain from saying things (uncivil remarks) which, without the new rule, they might want to say. But again, feedback effects, precipitated by the speakers' varying discursive dispositions – their varying appetites for risk, as well as their varying styles of discursive interaction – lead to what may have initially seemed like a surprising result, namely, more speech.

These are toy cases, but they help us grasp the crucial assumption/conflation that's involved in the standard account of how CEs work. On the standard account, instances of individual-level speech deterrence, when added-up, result in a more subdued discursive milieu. Individual Deterrence causes Group Suppression. But this doesn't account for the possibility of feedback effects like the ones in PEACEMAKER. If these effects are occurring, Individual Deterrence could unexpectedly increase the overall quantity of discussion.

III. Risk-Aversion, Intensity, and Heating

How would such feedback effects occur outside of the kind of toy scenario described above? In this section I describe a possible mechanism. The thesis, in essence, is that speakers who tend to be more risk-averse, and hence more susceptible to speech deterrence, can also tend towards moderation in what they say in a discussion. So, if a deterrent factor is introduced into a discursive milieu, more moderate speakers may be more deterred, meaning that the remaining speakers will tend to have more intense discursive interactions.

I explain and defend this hypothesis in what follows. One assumption in this, already indicated, is that people aren't uniformly risk-averse in their communicative dispositions. This is briefly noted in Schauer's account of CEs. Different people's "varying degree of risk aversion", he says, "will cause differing amounts of deterrence in situations where all other factors are the same".[20] This should be uncontroversial. It's conceivable that ordinary behavioral diversity could be less pronounced in the realm of communicative behavior. But absent any particular reason to think this is the case, Schauer's point seems to follow straightforwardly from the mundane fact – observable in everyday interactions, and in much behavioral research – that different people have different levels of risk-aversion across different contexts.[21] I'll be following Schauer in respect of this assumption.

---

[20] Schauer 1978, p. 698.

[21] Whether this is more innate or learned remains an open question. Shaw (1996) finds that risk-aversion is correlated with lower educational attainment. Cesarini et al. (2009) find that it is genetically inheritable. Recent work by Morgenroth et al. (2022) raises doubts about the long-standing thesis that women are more risk-averse than men. Classical prospect theory, made famous in Daniel Kahneman's work, utilises a generic model for risk-responsiveness, which doesn't model variability in risk-responsiveness across individuals. Still, the data that Kahneman and Amos Tversky used to evidence their theory, including in their ground-breaking article on the topic (1979), support my (straightforward) point: that different people evince variable degrees of risk-aversion, under different circumstances.

A. An Inverse Correlation Hypothesis

I'm going to coin a term to help analyse the effect I have just described. I am hypothesizing that there is an inverse correlation between Risk-Aversion and *Discursive Intensity*.

By Discursive Intensity (DI), I mean how vehement and bellicose people are in expressing their views. By *vehement* I mean confident about the rightness of one's views on contested topics. By *bellicose* I mean tending to disdain or dismiss other people's views, if they differ from one's own.[22] Speech is high-DI insofar as it evinces both sub-traits. By contrast, low-DI speech is more respectful and open-minded. The two sub-traits can be evinced either in speech's content, e.g. in saying that people's views are stupid, or immoral, or in speech's style and manner, e.g. in provocative or disputatious forms of address.[23]

Risk-Aversion (RA) refers to a person's tendency to prefer options with more certain outcomes over ones with less certain outcomes – as opposed to just trying to maximize gains. Given our topic, I'm interested in RA with respect to choices on when to speak or not speak, and what to say in speaking. High-RA speakers prefer speech acts (or omissions) with predictable results, whereas low-RA speakers are more blasé about the unpredictable consequences of their speech – e.g. consider people who vice signal, people motivated by totalizing ideologies, people who enjoy upsetting others (or playing the martyr), or speakers who simply don't feel bad about infringing rules and incurring penalties. Low-RA speakers like these are more willing to gamble on saying things that might go badly.

In practice, high-RA speakers will be reluctant to enter debates on controversial issues, because there is more uncertainty there about what speech (or non-speech) will provoke others and make one a target of resentment. Similarly, high-RA speakers will be warier about saying things that could violate speech restrictions – or which are at risk of being construed as violations – in a way that may lead to incurring costs or penalties. Low-RA

---

[22] Many studies construe *verbal aggression* as a trait that can be measured and correlated with other traits (e.g. cognitive flexibility) and behaviors (e.g. violence). For an overview of relevant literature see Rill et al. (2009). I haven't based my definition of DI on prior attempts to quantify verbal aggression; I want to leave it open what's the optimal way to conceptually frame an investigation of our hypotheses. My point here is that it isn't eccentric to posit an in-principle measurable behavioral trait or pattern along the lines of what I'm calling DI.

[23] For the sake of argument I'll assume that discursive vehemence and bellicosity go together for most people, most of the time. I believe this assumption is relatively uncontroversial, insofar as the kinds of personality types that conduce to discursive vehemence also conduce to bellicosity. (I'll say more about discursive 'trait-clusters' in Section III.C, discussing agreeableness.) Indeed, there's a kind of temperamental dissonance in a person's being bellicose *without* being vehement. (If I'm dismissive towards other people's views, doesn't this in some sense commit me to being confident about my views' correctness?) Then again, there's no dissonance, in principle, in being vehement without being bellicose. Being confident one is correct needn't dispose one to be dismissive of others' views. In any case, the phenomenon I'm positing is one where speech that's both bellicose *and* vehement becomes prevalent in a discursive context, as speakers who are neither bellicose *nor* vehement withdraw. My account doesn't make predictions about how Individual Deterrence affects group-level discourse in contexts where vehemence and bellicosity routinely come apart. But I invite the reader to follow me in the working assumption that these discursive traits commonly run together.

speakers, by contrast, will be more willing to join debates under either of these risky conditions.[24]

My hypothesis is that these two traits are inversely correlated. Naturally, the most plausible version of this hypothesis will build in some sensible qualifications.

First, the hypothesis needn't be that there's a very strong inverse correlation between the traits, so that high-RA strongly probabilizes low-DI. All I want to suggest is that higher RA probabilizes lower DI to some extent, and *vice versa*. So the pattern may look more like the loose correlation on the left in Figure 1, rather than the tighter correlation on the right.[25]

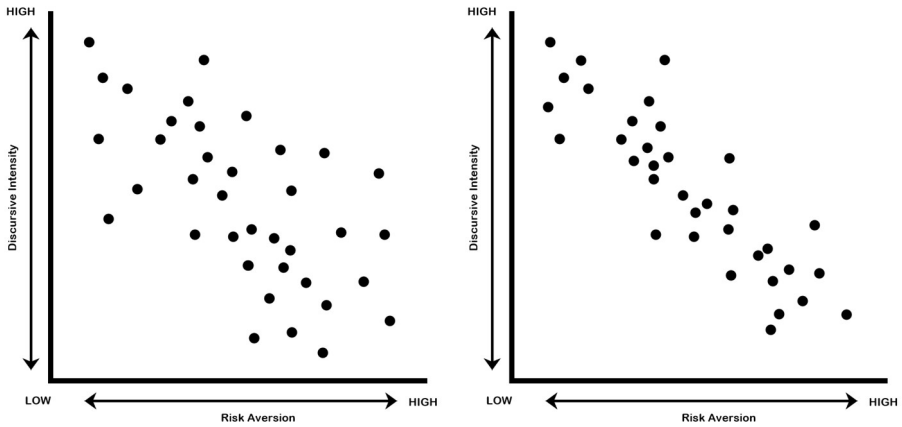

FIGURE 1: Schematic representations of (i) high-RA weakly probabilizing low-DI (left), versus (ii) high-RA strongly probabilizing low-DI (right)

Second, I don't want to claim that these traits have a high degree of stability within individuals across time, or across contexts. Someone who's high-I in one discursive context, like a social media debate about the topic of mental health, might be low-DI in another discursive context, like a water-cooler chat about politics. Schematically: someone who is high-DI, in some discursive context, C_1, at time t_1, may be low-DI in C_2 and/or at t_2.[26] The hypothesis I'm sketching isn't supposed to entail any controversial view about debated issues in social psychology, relating to the general stability or instability of people's

[24] I don't know of any research into how risk-aversion manifests in relation to communicative behavior in particular. As I say at the beginning of Section III, I'm assuming that ordinary variations in risk-averse behavior (whether more dispositionally- or circumstantially-driven) translate into people's communicative choices and behaviors.

[25] But is this sheer conjecture? Why think that such a correlation exists? To be discussed in Sections III.C and IV.

[26] For data indicating variability in risk appetite, see Soane and Chmiel (2005).

behavioral dispositions. The hypothesis is merely that these two behavioral traits manifest in a somewhat predictable pattern in relation to each other.[27]

So, the slightly refined hypothesis, incorporating these qualifications, is this: person A having high RA, in context C, at time t, probabilizes (to some degree) A having low DI, in C, at t. And conversely, A's having low RA probabilizes them having high DI (in C, at t).

B. Risk-Aversion and Discursive Intensification

How would an inverse correlation between RA and DI create an feedback effect, such that group-level discussion ends up being intensified by individual-level deterrence?

Consider PEACEMAKER again. The person who tends to mellow-out conflict, Pat, is also the one more susceptible to being deterred from speaking. Simultaneously, the people prone to firing up the discussion, Quinn and Rex, are less susceptible to speech deterrence. So, once a speech-disincentivizing rule is in play, Pat's withdrawal from the discursive arena allows Quinn's and Rex's latent discursive intensity to express itself to a greater degree.

If there is a non-trivial inverse correlation between RA and DI, in a particular context, C, then speech-deterring laws introduced in C will tend to generate a similar effect. We don't need to posit that all of the speakers who withdraw are, like Pat, self-conscious peacemakers. The main driver of the effect is an increased preponderance of relatively high-DI people, within some discursive group, aggravating and provoking each other more frequently, as lower-DI interlocutors withdraw, and each speaker's chances of conversing with a lower-DI interlocutor diminishes, in each discursive interaction.

In general – so this hypothesis goes – the speakers who are more susceptible to deterrence, in view of the potential risks / costs of speaking, and who are thus more likely to withdraw from a discussion – i.e. high-RA people – have lower DI. Conversely, actors who are relatively *less* susceptible to deterrence, and more likely to continue participating in discourse – those with lower RA – have higher DI. The introduction of a speech-deterring rule in C will naturally decrease the quantity of discursive participants. But at

---

[27] The situationist view of social psychology says either that character traits don't exist – that varied behaviors reflect responses to situational cues, rather than abiding traits, see e.g. Harman (1999) – or more modestly, that they don't have the stability needed to underpin neo-Aristotelian ethical theories that define right action in terms of virtuous character, see e.g. Doris (2002). In Section III.C, I present two factors that support my hypothesized correlation between DI and RA. One of these doesn't posit underlying traits. It suggests that discursively intense acts are inherently discursively risky, so that discursive risk-taking is basically *coextensive* with high-DI. The second does invoke traits that underpin the combination of high-RA and low-DI. But I don't commit myself to any claim about the stability of such traits. My thesis is that these two dispositions, RA and DI, manifest in a non-haphazard pattern. If someone manifests high-RA, in context C, this will tend to co-manifest with low-DI in C, and *vice versa*. This is compatible with the view that people exhibit high-RA in some contexts and low-RA in others. By analogy, to say that manifestations of shyness tend to be paired with manifestations of social anxiety, isn't to say there's a stable, cross-situational, combined trait (shyness-plus-social anxiety) that exists in some individuals but not in others. It's to claim that when person A's potential for social anxiety is elicited, in context C, A will also tend to manifest shyness in C.

the same time it will change the temperamental composition of C's pool of active interlocutors, so that there is a greater proportion of high-DI (vehement and bellicose) speakers, and a lesser proportion of low-DI speakers. It will change from being something like the left scatter plot, below, to something like the one on the right. Each dot represents a speaker. When a speech-deterring rule is introduced it leads to a withdrawal of higher-than-average RA actors, and this brings with it the withdrawal of lower-than-average DI speakers. So after the rule is enacted, the participant pool is made up of relatively more high-DI speakers, as in Figure 2.

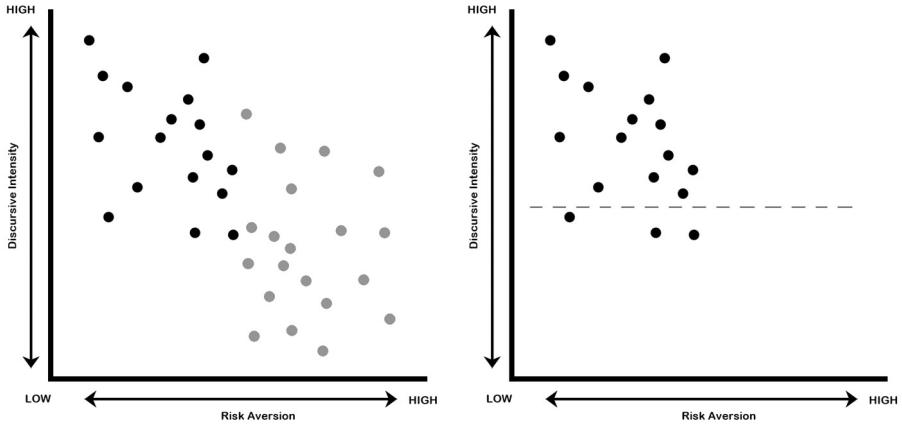

FIGURE 2: Withdrawal of higher-than-average RA speakers (highlighted left), entails a withdrawal of mostly low-DI speakers, so that the discursive pool is then left with an increased proportion of high-DI speakers (right)

Typically, if groups of more vehement and bellicose speakers are left to talk among themselves – without the tempering influence of comparatively sedate, open-minded interlocutors – the discursive climate gets intensified rather than subdued. This is an independent piece of conjecture, but it's fairly commonsensical and familiar from everyday experience. Vehement, bellicose speech, answered in kind with more vehement, bellicose speech, generates a discursive climate that mirrors the fractious temper of its participants, thus tending to increase both the quantity and affective intensity of communicative exchange.[28]

So, while it's true that fewer speakers will be participating in a discursive milieu which risk-averse speakers have withdrawn from, if the remaining participants tend to have higher DI, fewer speakers won't necessarily mean less speech, or more sedate speech. The volume and vehemence of discussion will be liable to increase under these conditions. An intervention that serves to deter individual speakers won't trigger a group-level

[28] We might expect this either because of some kind of Humean process of emotional contagion, or simply on the basis of mundane observations about what manners and modes of expression tend to escalate verbal conflict.

CE, in this event; what it is likely to lead to, at the group level, is some kind of HE. Individual Deterrence will result in group-level discursive intensification, instead of group-level suppression.

Discourse can of course be heated up via tactically provocative speech, like trolling. People can deliberately incite vehemence and bellicosity in their interlocutors, either as a means to some further end (e.g. as a 'dead cat' diversion), or as an end it itself.[29] Plausibly, we should be cautious about regulating public discourse in ways that make it easier for trolls and provocateurs to strategically foment conflict.[30] But notice: there needn't be any strategic provocation involved in the scenario where Individual Deterrence leads to an HE. The effect can be totally inadvertent. Like in PEACEMAKER, the aim of the rule that precipitates the effect may be reasonable, and everyone whose conduct is influenced by it may be reacting ingenuously. What causes the intensification, primarily, in the type of case that I'm positing, is the pattern of discursive behavior and risk-aversion across the group.

Granted, whether a HE occurs in any particular case will depend on fine-grained details of the scenario. To illustrate, consider a variant on the PEACEMAKER case, where Pat remains hyper-cautious about violating the pro-civility rule, but where, fortuitously, she receives clear advice about what the rule concretely permits, and reassurance that the rule will be applied in accordance with this advice. Thus reassured, risk-averse Pat may be confident enough to carry on in discussion, and any untoward effects of her withdrawal won't occur. We can imagine group-discussion scenarios in which, fortuitously, all of the Pat-like (i.e. low-DI + high-RA) speakers receive similar clarifications and reassurances about how to navigate the discursive risks. HEs won't materialise in such cases.

Moreover, sometimes any discursive intensification that does occur will be fairly mild, and unproblematic. The risks generated by the introduction of a speech-restrictive rule probabilize discursive intensification at the group level, but the degree of discursive intensification depends upon the strength of correlation. In short, the size of the effect, and the likelihood of it occurring, will both depend on how strong the correlation is between DI and RA in a particular context.

Still, the possibility of such cases – where HEs don't follow from Individual Deterrence, or where the effect is relatively mild – is compatible with the general pattern that I'm hypothesizing. In any context where the hypothesized inverse correlation between RA and DI obtains, the enactment of a speech-deterring rule will, other things being equal, tend to increase the proportion of vehement, bellicose speech, in the affected discourse, and correspondingly, increase the quantity of vehement, bellicose discursive interaction.

[29] On trolling, see Connolly (2021); on dead cat techniques, see Saunders (2019).

[30] Two examples of arguments along these lines are Simpson (2018) and Schauer (2020).

C. Support for the Hypothesis

Toy models aside, is there any reason to accept the pivotal hypothesis here – that a speaker, A, being high-RA (in context C, at time t) probabilizes A being low-DI (in C, at t)?

Here are two considerations that lend support to this correlation hypothesis.

First, there is, plausibly, an intrinsic connection between speech that's vehement and bellicose, and speech that's risky, in the sense of creating an unpredictable potential for negative reactions and blowback. If you arrogantly spout opinions, and tell other people their ideas are wrong or stupid, you are more likely to elicit anger, confusion, frustration, resentment, etc. Moreover, it becomes harder – compared to when you are speaking in a less vehement, bellicose way – to predict exactly what form other people's reactions will assume. After all, different people push back against disputatious speech in different ways. In short, high DI speech is volatile, in both senses of the word – tending to elicit reactions which are (i) relatively intense, but at the same time, (ii) hard to anticipate, in their particulars.

Why does this support the hypothesis that there is an inverse correlation between RA and DI? In short, it seems inherently unlikely that risk averse speakers will also speak in a discursively intense fashion, in a particular context, insofar as exhibitions of discursive intensity lead to risk and uncertainty in other people's reactions, in a way that directly strikes at the risk-averse speaker's aversions. That is, people who don't like risk will naturally tend to eschew conduct that increases risk. And high-DI conduct increases risk.

Second, there is, plausibly, a kind of synergy in being high-RA and low-DI. These are complementary traits that we will expect to see exhibited by people scoring highly in agreeableness, among the big five personality traits, i.e. people who, relative to the descriptive norm, tend toward behavior that's friendlier, more cooperative, and more respectful.

A number of studies indicate a correlation between agreeableness and risk-aversion.[31] One explanation of this is that being agreeable makes people more wary about upsetting others by actively courting risk (whereas, by contrast, for risk-friendly actors, lower agreeableness offers "insulation against guilt or anxiety about negative consequences").[32] As for DI, agreeableness is standardly explicated in terms of traits (e.g. altruism, gentleness, modesty), that can be understood to be either coextensive with lower-DI, or else conducive to it.[33] In short, someone who tends to get along well with

[31] The most consistent finding in research on the relationship between personality traits and risk-taking, showing up in studies that investigate the big five (or six) traits, is a positive correlation between risk-taking and openness. Low extraversion is also often found to be positively correlated with risk-taking. In any case, while it doesn't show up quite as frequently, there are still multiple studies finding a positive between agreeableness and risk-aversion, e.g. Nicholson et al. (2005), Hong and Paunonen (2009), Joseph and Zhang (2021); Salameh et al. (2022), and Ayers et al. (2023).

[32] Nicholson et al. 2005, p. 170.

[33] See e.g. Jensen-Campbell and Graziano (2001), Jensen-Campbell et al. (2003), and Sims (2017).

others is less likely to talk to people in a vehement and bellicose manner, and in addition, less likely to risk acting to bring about outcomes (via public discourse, or otherwise) which, as well as being bad for themselves, may cause interpersonal stress and drama.

Again, the hypothesis is that this correlation holds to some extent, in some cases – enough that there's some observable pattern in how DI and RA manifest in people's discursive conduct, rather than manifesting in a totally haphazard, unpredictable fashion.[34] Whether this correlation results in notable group-level discursive effects will also depend on the group's initial temperamental composition. If a discursive context initially has a low proportion of low-DI / high-RA agents, an intervention that leads to those agents' discursive withdrawal will make less difference to the group's composition, and is therefore less likely to intensify discourse. This means Individual Deterrence is unlikely to leads to HEs in contexts where speech-related risks are generally mild – like, say, an anonymous online forum. Individual Deterrence is more likely to lead to HEs in contexts where people are held accountable for (actual or perceived) discursive misconduct – like, say, university campuses, or the internal communication channels of large media organizations.

While I've cited some supporting evidence, this is obviously still conjectural. But the conjectural nature of the thesis fits fine with my overall argument. I'm trying to show that discussion of CEs is premised on a dubious assumption: that Individual Deterrence naturally leads to Group Suppression. This assumption only holds given other assumptions vis-à-vis the causal relations between individual-level behavioral traits and group-level discursive phenomena – assumptions which, in extant work on CEs, are barely recognized, much less backed by evidence. The point of my analysis is to suggest that Individual Deterrence and Group Suppression don't necessarily go together. The burden of argument isn't to irrefutably verify Section III.A's inverse correlation hypothesis. What has to be shown is that there's some reason to think this correlation obtains to some degree. The hypothesis needs to stand as a credible one. I've tried to show that it does. We cannot assume that Individual Deterrence always (or typically) leads to Group Suppression, because this ignores the credible possibility that people's susceptibility to speech deterrence exhibits a patterned relationship to the types of expressive contributions that people are disposed to make.

IV. Hate Speech and Heating

In this section I offer further argumentative support for the account in Section III, by showing how it helps to explain a puzzling pair of observations related to anti-hate

[34] As per my comments in note 27, I'm uncommitted on the cross-contextual stability of these traits. The hypothesis is that RA and DI tend to be exhibited in a predictable pattern across situations. My point in the above is that if we *do* think personality traits are stable enough to figure in our hypotheses, then there's a plausible explanation available about why possession of one trait, agreeableness, would elicit both high RA and low DI, as per the inverse correlation hypothesis from Section III.A. Still, the untenability of personality type theories (if such theories are untenable) shouldn't cast significant doubt this hypothesis's plausibility. It's a thesis about how certain behaviors are exhibited in a patterned way, not about exactly what causally underpins that pattern.

speech law. These laws seem to have a non-trivial deterrent effect on individual expression. But they don't seem to lead to a general stifling of public discourse – quite the opposite. My proposal is that an HE may be occurring in cases where these two phenomena co-exist.

Let's set this in context. CE-related worries are often expressed in debates around anti-hate speech law and other policies restricting discriminatory speech. Some Liberals believe that these restrictions lead to self-censorship, especially among conservative or moderate speakers.[35] People worry that they could incur penalties due to overzealously applied restrictions, or simply that costly allegations of a breach may be brought against them. The danger posed by anti-hate speech laws, then, according to Gerard Anderson, is

> an insidious chilling of political debate, as people censor themselves in order to avoid legal charges and the stigma and expense they bring. And the most serious chill is not of fringe racists but of mainstream moderates and conservatives.[36]

Anderson believes this suppression is mainly due to "uncomfortable and expensive brushes with speech laws".[37] Expanding on these concerns, Nadine Strossen argues that even when anti-hate speech laws are narrow in scope, they still "repose great discretionary power in enforcing officials", and where these powers are given, she says, officials "consistently have exercised [them] to suppress unpopular views", in a way that has "chilled yet more expression, including mainstream political views".[38] Strossen says that these laws also inhibit the kind of intergroup dialogue that helps to mitigate social conflict. Thus, she says, they have "a chilling impact on both open expression and open-minded listening".[39]

Similar claims are found in parallel debates about other policies that, like general anti-hate speech restrictions, aim to curb discriminatory speech. Consider debates over the so-called *Working Definition of Antisemitism* (WDA), promoted by the International Holocaust Remembrance Alliance. The WDA has been incorporated into codes of conduct and thus become a part of speech-restrictive policy in some US and UK institutions.[40] Critics say that the WDA's adoption in speech codes muddies the distinction

---

[35] For the sake of argument we can grant the pejorative notion of self-censorship that's assumed here, and bracket the broader question of how to distinguish between pernicious self-*repression* and healthy/valuable self-restraint. For discussion of these questions, see e.g. Horton (2011) and Festenstein (2018).

[36] Alexander 2006.

[37] Ibid.

[38] Strossen 2018, p. 104.

[39] Ibid, p. 150. Other examples of authors voicing concerns about CEs triggered by anti-hate speech laws, in the literature on hate speech, include Wolfson (1997), Appiah (2012), and Heinze (2016).

[40] In the US, for instance, a 2019 presidential executive order, directed to agencies enforcing Title VI of the Civil Rights Act, mandated that the WDA and its accompanying examples be used as evidence of discriminatory intent in investigation allegations of anti-Semitic incidents in federal institutions. This was widely interpreted as an attempt to suppress criticism of Israel in American universities; see e.g. "Trump targets antisemitism and Israeli

between antisemitic speech and legitimate criticism of Israel, and thus deters the latter.[41] Granted, there's a penumbra of uncertainty that surrounds any identity-protective speech rule. One can never be 100% certain, under such rules, that one's permissible mention of an identity-prejudicial idea will not be misjudged and penalized, by a confused adjudicator, as a use or endorsement of that idea. But the critics' objections to the WDA go further than this. They worry that by overtly complicating the distinction between antisemitic speech and criticism of Israel – by stating that the latter can constitute the former, in certain contexts and cases – the WDA prevents the kind of *ex ante* clarifications that would be needed to reassure risk-averse speakers that their lawful criticisms of Israel will not be adjudged antisemitic. Indeed, one of the WDA's authors has claimed that this is the motivation behind its being institutionalized, as a tool for assessing allegations of antisemitism in universities – to generate a CE around criticism of Zionist viewpoints and Israel's military actions.[42]

It is easy for supporters of restrictions on discriminatory speech to deride such concerns.

> Discriminatory and bigoted ideas are being suppressed and discouraged, you say? You could have fooled me! Surely it is *easier*, today, than it has been at any point in the last few decades, to express discriminatory and bigoted attitudes in public.[43]

In short, if anti-hate speech laws and the like were triggering CEs, then we should expect to observe a stifling or suppression of debate on topics related to the kinds of speech that these restrictions affect, e.g. debates about race, religion, and nationalism. We should expect to find racists and antisemites self-censoring, and a withering of debate around the kind of controversies where those people tend to speak. But that seemingly isn't what we find. The debates and slanging matches don't appear to be stifled at all; they seem relentless and spirited. So the idea that speech restrictions are causing CEs, in discourse adjacent to identity-based discrimination, seems borderline neurotic. Worries about the over-deterrence of this speech is, so one might argue, an ideologically-driven overreaction to anecdata.[44]

---

boycotts on college campuses", *New York Times*, 10th December 2019, www.nytimes.com/2019/12/10/us/politics/trump-antisemitism-executive-order.html.

[41] For arguments to this effect, which mention CEs, and which argue that, despite being touted as a non-legally binding definition by the UK government, the government's 'adoption' of the WDA has afforded it a quasi-legal status, in how it affects discourse on Israel in the UK, see Gould (2022), Deckers and Coulter (2022).

[42] Stern 2019.

[43] Consider the following remarks by Malik (2019), on 'the myth of the free speech crisis'. This myth's purpose isn't to protect free speech, she says, i.e. "the right to express one's opinions without censorship, restraint or legal penalty". Rather, its aim is to "normalise hate speech", to "shut down legitimate responses to it," and to "secure the licence to speak with impunity". The idea that it's easier today to publically express racism is often linked to the rise of reactionary populism in electoral politics – the rough thesis being that figureheads are simultaneously gaining support from and emboldening 'grassroots' racist attitudes. For an analysis of these feedback loops in the rise of today's reactionary political leaders, see Jacobs and van Spanje (2020).

[44] One example of this is Bedi's effort to debunk claims about CEs resulting from content moderation on social media platforms. "Rather than necessarily being chilled in the traditional sense," he says, the data suggest that "social media users can communicate what they want even with restrictions, and when they are restricted, they communicate in a more civil manner"; see Bedi (2021, p. 305). An example of this view in mainstream debate is a

Granted, this reasoning relies on a counterfactual claim that's hard to establish. Consider criticism of Israel, and the impact of anti-antisemitic speech codes involving the WDA. It's hard to know just how much criticism of Israel there would be without the influence of the speech codes that are allegedly deterring it. Anecdotal evidence can be offered on both sides – evidence of people feeling muzzled by these codes, but also, of people vigorously rebuking Israel in spite of them.[45] The question is: what patterns (if any) do the anecdotal data indicate? If I think criticism of Israel is being chilled, I might say: "who knows how much *more* criticism there would be, but for these CEs which are being elicited by the WDA?

However, someone who doubts the existence of CEs in this area can give a quick reply. On a standard account of how CEs work, deterrence is greatest for speech that's most at risk of penalty. For example, anti-defamation laws deter serious defamation far more than *borderline* defamatory speech. This is due to the basic prudential logic of risk-avoidance. Borderline defamation is less likely to be met with a lawsuit, or to be penalized in the event that it is. So people are only mildly deterred from saying borderline defamatory things. As Schauer says, it's where speech "falls close to the line separating protected and unprotected" that it is "most likely to be erroneously adjudged unlawful." So "the degree of fear", and the CE's deterrent impact, "will be greatest where such borderline activities are involved."[46]

Drawing on this, we can offer a plausible prediction about how a discursive milieu will look if it's being affected by a standard CE. If anti-defamation laws are deterring lawful speech, we won't see a relative abundance of seriously defamatory speech (e.g. allegations of criminal wrongdoing) relative to borderline defamation (e.g. allegations of lawful immorality). We expect to see relatively more borderline defamation than serious defamation.

A corresponding prediction can be applied to anti-hate speech laws and the like. If these are deterring a significant amount of lawful speech, e.g. about identity-political topics, then – given a standard analysis of how CEs work – we shouldn't expect to find a relative abundance of highly controversial / risky speech on these topics. That speech will be more susceptible to deterrence, as speakers try to avoid the risk of incurring costs linked

piece called "A more specific letter on justice and open debate" (objectivejournalism.org/2020/07/a-more-specific-letter-on-justice-and-open-debate/), from 2020, by a group of writers responding to 'the Harper's Letter' (i.e. "A letter on justice and open debate"; harpers.org/a-letter-on-justice-and-open-debate/). The reply-letter's authors run through a series of evils which, according to the Harper's Letter, result from 'cancellation' practices in the media, e.g. editors being fired, books being withdrawn, topics being declared off-limits, etc. In each case, the reply argues that these allegations are an exaggerated portrayal of isolated incidents. In a similar vein, commenting on the WDA, Nelson (2021) says "fear that there will be a chilling effect on anti-Zionist speech on… campuses has not been borne out by reality… Although some Jewish groups have called for the suppression of certain forms of anti-Zionist speech… they have not prevailed. Similarly, NGOs of many stripes routinely call on universities to censure or fire faculty for remarks of all kinds, but universities routinely dismiss those demands, except for part time or contingent faculty".

[45] See the references in notes 40-42 for discussion of numerous anecdotal examples, lending *prima facie* support to either side of this controversy.

[46] Schauer 1978, p. 696.

to actual or perceived rule-breaking. We will expect to see a relatively greater incidence of mildly controversial speech, in these areas, compared to highly controversial speech.[47]

This expectation seems to be confounded in today's Liberal societies. On issues connected to race, religion, and nationality, on which lawfully expressible views are supposedly being self-censored, due to CEs triggered by speech-restrictive rules, public discourse isn't characterized by a relative abundance of cool, sedate, non-provocative speech. Oversimplifying a bit – and granting that availability heuristics may be coloring our perceptions – discourse on these topics seems to involve a lot of vehemence and bellicosity. It isn't characterized by many people speaking sensitively and open-mindedly.[48]

Many causal factors may be appealed to as part of an overall explanation of this. Indeed, it's at least possible that there are two independent, parallel phenomena at work here. In some discursive contexts, speech-restrictive rules have led to Individual Deterrence; at the same time, alternative discursive spaces have opened up (e.g. online), where vehement and bellicose speech is encouraged or amplified.[49] (But notice: this alternative hypothesis wouldn't be well-positioned to explain why Individual Deterrence coexists with group-level discursive intensification in older, offline discursive contexts, like universities, or in legacy media organisations.) In any case, for our purposes, the crucial point is that this pattern – namely, the apparently widespread coexistence of Individual Deterrence, with group-level discursive intensification – isn't what an orthodox account of CEs predicts.[50] So the claim that an individual-level chilling of moderate opinion is leading to some sort of broader suppression or stifling of group-level discourse still seems relatively dubious.

My account of HEs in Section III offers us another way of interpreting what's going on. The idea is that a significant number of speakers *are* being deterred from engaging in

---

[47] More precisely: we should expect to see more mildly controversial speech and less highly controversial speech, *relative to the assumed baseline incidence of these kinds of speech*. If the assumed baseline is that *most* people have super-controversial views on issues linked to race, religion, and nationality, then we wouldn't necessarily expect highly controversial speech to be less prevalent overall, even if it is more deterred that mildly controversial speech.

[48] Thinkpieces and op-eds centred on this observation abound. In one emblematic instance of this genre, a 2022 *New York Times* editorial opened by declaring that "Lately, everyone seems to be mad – all the time… some crucial layer of emotional regulation has disappeared. It's as if our collective gears have been stripped by the isolation and unspooling of the last few years"; "The year we lost it", 17th December 2022, nytimes.com/interactive/2022/12/17/style/ 2022-year-of-rage.html. Systematic longitudinal data on the (apparent) increasing prevalence of combative and controversial speech in public discourse – data that go beyond geographically and temporally localized survey data – are harder to find.

[49] For example, the prevalence of combative speech on topics linked to identity-injustice may be attributed to a (causally distinct) increased polarization in liberal societies, accelerated by social media echo chambers. These phenomena have been examined in various popular social science books in recent years, e.g. Mason (2018) and Bail (2021). Recent work suggests that a factor driving increased hostility on social media is the greater visibility of bellicose speakers. Roughly, social media amplifies bellicosity, rather than causing or attracting it; see Bor and Peterson (2022). Another *prima facie* (partial) explanation, of why there's more combative speech on topics related to identity-injustice, today, is that there are greater economic incentives nowadays for people to produce such speech; see Williams (2023).

[50] At any rate, not unless we start with tenuous baseline assumptions (see note 47).

debates on issues that anti-hate speech laws are connected to, but that this isn't stifling the debate, as a standard CE analysis predicts – rather, that it's causing an intensification of debate. As Anderson suggests, it may be, mostly, people with milder views who are self-censoring – people who, in keeping with their sedate tempers, also tend to be more risk-averse about joining discussions in which penalties for ill-considered remarks might await them. But with those people withdrawing from the discussion, there is increased volatility in the temperamental composition of the discussant pool, which in turns tends to intensify debates.

Granted, things won't be as neat and tidy as this sounds, in messy real-world cases. Which speakers self-censor, in a discursive context, will depend on a range of circumstantial specifics. But for HEs to occur, it doesn't need to be the case that all moderate speakers self-censor. Only a good portion of them – and greater portion relative to the higher-DI speakers – need to self-censor, in order for some kind of HE to ensue.

The standard account of CEs tells us that Individual Deterrence causes Group Suppression. This account leaves space for two big-picture interpretations of what's going on in today's liberal societies, in relation to identity-protective speech laws. Interpretation 1: maybe appearances are misleading, and there actually aren't a lot of people being deterred from entering discussions about controversial topics linked to identity-prejudice. Or, Interpretation 2: maybe a lot of Individual Deterrence *is* occurring, and is stifling debate in these areas, but we're getting a misleading impression that the debate isn't being stifled. What I'm offering is a wholesale alternative interpretation, that lies outside the horizons of the standard account of CEs. What's going on is not a CE, but rather a HE. Individual Deterrence is occurring, at a non-trivial scale, but this is altering the temperamental profile of public discourse, in a way that's making affected debates hotter rather than cooler.

V. Why is Speech Deterrence Wrong?

Let's shift focus. Why is the inadvertent deterrence of lawful speech wrong or bad? We talk about CEs because we think we should try to prevent or mitigate them. But why? Why are they bad, over and above the generic badness of having imprecise laws that we're uncertain how to abide by? One answer, popularized in Schauer's account, is that speech is transcendentally valuable. But this seems too sectarian to compel wide support. In this section I'll draw upon my account of HEs to propose an alternative answer. Inadvertent deterrence of lawful speech is objectionable because it conduces to dysfunctional public discourse.

A. Transcendental Value

To begin, it's helpful to understand why Schauer appeals to speech's transcendental value in seeking to explain the problematic character of speech deterrence.

Inadvertent deterrence can result from any conduct-limiting rule. For example, people may be deterred from engaging in borderline-fraudulent activity due to uncertainty about the legal boundaries of fraudulence. Governments have Fullerian reasons to limit this kind of 'fuzzy-borders' deterrence, i.e. reasons reminiscent of Lon Fuller's account of law's internal morality. Against H. L. A. Hart's conceptual decoupling of a rule's moral and legal status, Fuller argues that rules can't be *bona fide* laws unless they fulfil various rule-of-law-related normative standards, such as predictability, publicity, and prospectivity.[51] The basic idea is that it's antithetical to legal governance to put people in a scenario where they're uncertain how to abide by the rules that govern them. This makes people more vulnerable to authority's caprices. We need to be able to foresee the legal ramifications of our choices, and deliberate accordingly, which is hard if laws are erratic, retrospective, etc. And whether or not one agrees with Fuller that a failure to meet these standards nullifies a putative law's legality, these standards are surely appropriate regulative ideals for law. Law should alleviate the burden of trying to guess how our decisions may go awry due to authority's whims.

This burden will be exacerbated by speech restrictions that generate uncertainty about what is lawfully sayable, and thereby deter lawful speech. But the badness of that, as indicated in the fraud comparison, isn't in any way distinctive. If the remedy for CEs is that "lawmakers should seek to minimise… uncertainty in the clear design of foreseeable and accessible legislation",[52] then our remedy is merely a prophylactic for all law-making. We haven't yet hit upon a reason to take special care in how we design *speech*-restrictive laws. If there's something distinctively bad about inadvertent speech deterrence, we ought to be able to say why this deterrence is different to (presumably, worse than) the deterrence of other acts.

This is the explanatory target that Schauer is trying to strike. He's trying to explain why we might think "an erroneous limitation of speech" – a rule under which speech that shouldn't be limited, *is* limited – has "more social disutility than an erroneous overextension of freedom of speech".[53] If we want to avoid this greater disutility, we need an approach to speech regulation that errs in a permissive direction. This means permitting speech which, on its own merits, may warrant restriction – like with the actual malice rule in US defamation law (see Section I). Erring against over-deterrence of speech makes sense if we think that the deterrence of good speech is much worse than the non-deterrence of bad speech – a bit like Blackstone's Ratio in criminal law, on which it's better for ten guilty people to go free than for one innocent to be punished.[54] Schauer's point is simply this: if "the transcendent value of speech receives the same priority that Blackstone gave to individual liberty", then we have a distinctive objection to speech

[51] Hart (1958); Fuller (1958).

[52] Townend 2017, p 80.

[53] Schauer 1978, p. 688.

[54] Here I'm paraphrasing Schauer 1978, p. 708.

being over-deterred.[55] The objection isn't rooted in a generic Fullerian ideal, related to the rule of law, but rather, a narrower Blackstonian thesis about the special disvalue of deterring a particular type of activity – namely, speech.

B. The Sectarianism Issue

This resolves the distinctiveness issue. But it replaces it with a sectarianism issue.

We want to know what makes the deterrence of lawful speech a bad thing not just for followers of particular moral viewpoints, but for all reasonable people with a stake in law and policy. If the explanation we provide to this end is that speech has transcendent value, this desideratum isn't fulfilled. Schauer links his account of CEs to classical Millian ideas, suggesting that we prioritize free speech above other goods and ideals "because of the overall societal benefit that is presumed to flow from the uninhibited exercise of first amendment freedoms".[56] This isn't an absurd presumption. But nor is it an obvious truth that all reasonable people endorse or should endorse. It's a sectarian thesis that many reasonable people reject. Reasonable people can deny that speech is superordinately valuable, as Mill and his followers believe, or that its value, however great, is greater than other valuable things. Even plenty of liberals are reluctant to exalt speech's value in this way.[57] To object to speech deterrence on this basis commits us to what Joshua Cohen calls a Maximalist theory of free speech, on which the downsides of free speech, such as they are, are trumped by speech's overriding and exceptional value.[58] This Maximalism is, at minimum, in tension with the pluralistic and anti-perfectionistic leanings of contemporary liberal theory.[59]

Here is a less sectarian proposal. Deterrence of lawful speech is bad because it conduces to dysfunctional public discourse, and this undermines a good – namely, stable and cooperative governance – whose value can be affirmed by all reasonable conceptions of the good. When HEs occur, and more moderate, open-minded speakers withdraw from

---

[55] Ibid, p. 704.

[56] Ibid, p. 691.

[57] E.g. both Brink (2001) and Waldron (2012) defend anti-hate speech laws – and criticize the notion that free speech overrides *pro tanto* justifications for such laws – within the parameters of a forthrightly liberal conception of justice.

[58] Cohen 1993, p. 220.

[59] Of course I'm not making any headway, via these cursory remarks, in debates between perfectionist and political liberals. A viable perfectionistic liberalism may be in the offing, as far as anything I'm saying here goes. But note that even the most influential defense of perfectionist liberalism in modern political philosophy, i.e. Raz's autonomy-based perfectionism, in *The Morality of Freedom* (1988), faces a serious challenge in providing a principled – as opposed to merely *ad hoc* – explanation about why it's wrong to paternalize people to prevent autonomy-impairing behavior; see Quong (2011, chapters 2-3). And the form of perfectionist liberalism we're contemplating above, i.e. one invoking speech's alleged transcendental value, is much less attractive than one that's grounded in some putatively indispensable ideal of autonomy. In short, even if one believes there are ways to defend perfectionism as a preferable interpretation of the liberal tradition, a perfectionism that's grounded in such an overtly sectarian ideal, as speech's transcendental value, seems untenable.

public discourse, debate about important subjects is more likely to become a war of words between polarized factions. And this makes it harder for public discourse to play the role we reasonably want it to play, in informing and guiding our collective governance decisions.

So, suppose we're assessing a policy measure whose proximate aim is to mitigate the inadvertent over-deterrence of speech, e.g. some clarifying caveat on our anti-hate speech laws, which spells out the difference between religious critique and religious vilification. Suppose we're asked what our justification is, for trying to mitigate this over-deterrence? If our justification here cites this measure's utility in arresting a slide towards dysfunctional public discourse, this seems to satisfy the anti-sectarian justificatory demands laid out in mature public reason liberalism, of the kind made prominent by Gerald Gaus among others, and whose core ethical concerns are shared by most post-Rawlsian political liberals.[60]

My point here should make sense irrespective of how it's situated in relation to turf wars in liberal theory. Most people defending free speech don't want to base their arguments upon idiosyncratic values. They want to tell a broadly-appealing story, about why protecting free speech is the sensible and just thing for societies to do. In philosophical work on free speech, beyond the narrow issue of how we analyse CEs, the most compelling free speech justifications are those that appeal to goods and ideals which most reasonable people endorse – not eccentric theses about speech's transcendent specialness, but more modest theses about the value of cooperative communication and genuine engagement with alternative viewpoints. By shifting focus away from parochial conjectures about speech's transcendental value, and emphasizing the destabilizing potential of dysfunctional public discourse, we can integrate our intuitive (but, thus far, under-theorized) worries about the inadvertent deterrence of lawful speech, with the kind of moderate and broadly-appealing normative premises that show up in the most compelling free speech arguments.[61]

In sum, if inadvertent speech deterrence is bad because it conduces to dysfunctional public discourse, then we have a justification for measures aimed at mitigating inadvertent deterrence that is both distinctive, i.e. it doesn't reduce to a Fullerian ideal vis-à-vis law's deliberative utility, but also appropriately non-sectarian, i.e. it's cognizable not just

[60] See in particular Gaus (2010).

[61] Among classic works in free speech theory, I have in mind especially Meiklejohn's *Free Speech and its Relation to Self-Government* (1948), and its envisioning of free-speech-protected public discourse as something akin to a town hall meeting where we discuss our concerns in a process of collective self-government. The ideals Meiklejohn invokes, in this portrayal of free speech's foundations, don't seem like a front for some parochial conception of the good. Meiklejohnian ideals make a claim on us because of a fact of political life that all conceptions of the good have to reckon with – that life must, on pain of Hobbesian chaos, be lived in minimally cooperative coordination with others.

for hard-line libertarian speech-lovers, but under most reasonable conceptions of the good.[62]

C. Chilling, Heating, and Dysfunctionality

What is the link to HEs? Under a standard account of CEs it isn't evident why Individual Deterrence of lawful speech – and the Group Suppression that this purportedly leads to – adds up to any discursive dysfunctionality. Think of it like this. Suppose we enact an anti-hate speech law which, as well as deterring harmful discriminatory speech, as it's intended to, also deters some non-discriminatory and relatively benign speech about nearby topics – speech which is still potentially liable to be misjudged as harmful. This deterrence is *prima facie* regrettable. But unless it becomes so widespread that public debate totally fizzles out, it needn't have a significant impact upon our society's ability to host robust, wide-ranging debates on topically-adjacent issues. Simply put: a less overcrowded discursive participant pool does not, by itself, make public discourse dysfunctional. As Alexander Meiklejohn famously stated, healthy public debate "does not require that, on every occasion, every citizen shall take part… what is essential is not that everyone shall speak, but that everything worth saying shall be said."[63] In short, CEs, as they are ordinarily understood, will not necessarily have a significant adverse impact on the overall well-being of public debate.[64]

But if Individual Deterrence of lawful speech leads to multiple kinds of disruptive group-level discursive effects – not just stifling CEs, but also intensifying HEs – then it is easier to see how and why this deterrence will typically result in dysfunctional public discourse overall. Again, suppose that a newly-enacted law is deterring some genuinely harmful discriminatory speech, as well as some relatively benign speech that is merely liable to be perceived as harmful. Even if this deterrence isn't widespread enough to totally stifle topically-adjacent debates, it can still inhibit a society's ability to conduct cooperative and robust discussions about topically-adjacent issues, by infelicitously altering the temperamental composition of the discussant pool, as per my account of HEs in Section III. Any HEs that are triggered in this way will tend to exacerbate the hostility and mistrust which characterize much public debate in contemporary liberal societies, as moderate

---

[62] Granted, there are sectarian ways of valuing respectful public discourse, i.e. defenses grounded in a perfectionistic ideal of civility. But recent philosophical defenses of civility, e.g. Bejan (2017) and Olberding (2019), emphasize civility's connection to the legitimation requirements for governance in diverse societies, and thus they align with my argument's more 'public-reason-liberalism-friendly' understanding of civility's importance.

[63] Meiklejohn 1948, p. 25.

[64] Granted, here I'm rejecting a key thesis in *On Liberty*, where Mill posits a link between clashes of opinion and the vitality with which opinions are believed. For Mill, all suppression of public debate leads to discursive dysfunctionality, by inhibiting the clashes of opinion that prevent opinions from becoming 'dead dogmas'. But this part of Mill's argument is unusable, for present purposes, because its operative notion of *vitality* is deeply parochial, such that its prescriptions, while formally utilitarian, are sectarian in substance; see Gray (1991, pp. xxv-xxx).

voices with higher RA withdraw from the discursive arena, and as the temper of the conversation comes to reflect the more vehement and bellicose traits of the lower RA speakers who carry on participating.[65]

Some authors who worry about CEs, like Strossen, oppose most restrictions on discriminatory speech.[66] But as noted in Section I, there is also a more moderate policy stance on this issue, among liberals, which says that restrictions should be limited in the forms of speech they apply to, and buttressed with guidelines that offer reassurance to actors who may be worried about inadvertently running afoul of the constraints – something akin to the actual malice rule in US defamation law.[67] This applies even if we think there are good reasons for limiting discriminatory speech. Even if well-known claims about the harmfulness of hate speech are correct, and even granting that free speech rights are not infinitely stringent, the careless deterrence of lawful speech can affect a society's ability to discern and discuss the issues of the day. Our goal of preventing harm shouldn't lead us to enact restrictions that end up deterring speech that's contentious but still lawful and relatively benign.

My account of HEs (in Sections III through IV) combined with my account of why Individual Deterrence is objectionable (in this section) shows why we don't have to be free speech fundamentalists, with a fetishistic notion of speech's superordinate value, in order to see things this way – to think that the inadvertent deterrence of lawful speech is something we should mitigate via suitable policy measures. Inadvertent deterrence of lawful speech doesn't necessarily result in CEs. Sometimes its group-level impact, such as it is, is to trigger HEs, and thus to exacerbate the dysfunctionality of public debate on

---

[65] In positing a link between (i) dysfunctional discourse, and (ii) bellicose discourse, am I condemning anger in public discourse? No. Expressions of anger needn't be bellicose in the relevant sense. Bellicosity means a tendency to deride other viewpoints. But fairly deriding a view that wholly merits derision isn't bellicose. There must be some indiscriminate ascription of immorality or stupidity in play. HE-affected discourse is dysfunctional not because it features expressions of anger, but because it's rife with indiscriminate derision. Moreover, as Srinivasan (2018) argues, *affective injustice* in public discourse owes to the fact that anger at injustice may be apt but also counterproductive in rectifying the injustice occasioning it. Under non-ideal conditions, mitigating affective injustice requires societies to react justly to apt anger, so that it doesn't so easily backfire. Anti-HE policies can be adjudged favorably, by this measure. Their point isn't to exclude anger (apt or otherwise) from public discourse, but to counteract the (self-)exclusion of non-bellicose expression. In a healthy discursive ecosystem, where the non-bellicose haven't pre-emptively withdrawn, expressions of apt anger are less likely to lead to dysfunctional or counterproductive results. An anti-HE policy agenda seeks no support from the kind of anti-anger politics that Srinivasan critiques – the kind that tells victims of injustice they must mute their anger at injustice, for the sake of peace and welfare moving forward.

[66] "Even if a 'hate speech' law were written relatively narrowly," Strossen says, "it would be 'the worst of both worlds'; due to its inherent vagueness, it still would repose great discretionary power in enforcing officials" (2018, p. 104).

[67] For example, consider the following from Blackford, concluding his discussion of the perils of speech restrictions. "The moral of this story," he says, "is not that the state must totally keep out of regulating speech that involves religious, racial, cultural, and similar sensitivities… some of this speech is grounded in forms of hostility that can rise in intensity to the worst kinds of racism… [But] any restrictions on speech must be scrutinized constantly. This includes the way the laws are drafted and the way they're interpreted and applied"; see Blackford (2019, p. 77). Brown makes a similar point (2015, pp. 267-268).

important issues. This effect is bad news for everyone, not just for the free speech fanatics or über-libertarians in our midst.

VI. Conclusion

Bertrand Russell said "in the modern world the stupid are cocksure, while the intelligent are full of doubt".[68] I have argued that our understanding of inadvertent speech deterrence – of the phenomena we usually refer to as Chilling Effects – should be informed by a similar hypothesis. People are not uniformly risk-averse. If more risk-averse actors also tend to be more moderate in the views that they express, then a restriction which deters lawful speech will also affect the temperamental composition of the pool of speakers participating in public discourse. Much like the situation that Russell is adverting to – in which "the intelligent are full of doubt" – valuable contributors will withdraw from the discussion, and the discussion will deteriorate as a result. But in the cases that I have been discussing, this deterioration will be crucially unlike what's envisioned in standard accounts of CEs. It will involve an intensification of debate, rather than a suppression or stifling of it.

The HE hypothesis, and the inverse correlation hypothesis that underpins it, are only hypotheses. But I have tried to show that they are credible hypotheses. I offered two arguments to this effect in Section III.C, and in Section IV I explained how these hypotheses provide a plausible explanation of a puzzling pair of observations. Anecdotally, restrictions on hate speech do appear to deter some lawful speech related to issues around identity-based injustice. But it seems implausible, given the vigor and relentlessness of debate around those issues, that these restrictions are in any significant way stifling public discourse in this domain. If hate speech restrictions and the like are triggering HEs, instead of standardly-analysed CEs, then these two observations actually make pretty good sense, side-by-side.

The thing to do with promising hypotheses is to search for evidence that sheds further light on their truth or falsity. The data might not end up favoring my hypotheses about HEs, and challenges may arise in observing and quantifying the traits that I've been describing under the label of Discursive Intensity, or in designing experiments in which the patterns of discursive aversion and enthusiasm that seem to occur in real-world debate are elicited. People know that they won't be fired or hit with a costly lawsuit for making badly-received remarks during a lab experiment. This could place a qualifying asterisk next to any lab-based data concerning the relation between Discursive Intensity and Risk-Aversion.

But however the chips fall there, there are still two significant takeaways from all this, if we ascribe a non-trivial credence to Section III's correlation hypothesis – as I think we should.

---

[68] Russell 1998, p. 28.

First, as I argued in Section V, this hypothesis helps us see why we still have good reasons to try to mitigate the inadvertent over-deterrence of speech, regardless of whether we regard speech as transcendentally valuable, in the way that some gung-ho libertarians do.

Second, as I argued in Sections II through IV, existing discussions of CEs mistakenly lead us to view Group Suppression as a natural or inevitable result of Individual Deterrence. We are interested in Individual Deterrence because it seems likely to ramify out into group-level discursive effects. However, the nature of those group-level effects depends on precisely how individual risk-aversion is related to people's discursive behaviors and temperaments. Most discussion of CEs has taken for granted controversial theses about the relations between individual-level discursive traits and group-level discursive phenomena – theses that typically haven't been acknowledged, much less critically dissected, or backed up with evidence. My arguments show us why any normative analysis of CEs needs to distinguish the individual- and group-level phenomena that existing accounts have bundled together.

Acknowledgements

Thanks to the editors of this journal and two anonymous referees for feedback on an earlier draft. This work also benefited from the input and criticism of Zachary Berenbaum, Quassim Cassam, Adam Dean, Rachel Fraser, Jonathan Gingerich, Toby Handfield, Daniel Hemel, Jeff Howard, Leslie Kendrick, Maxime Lepoutre, Neil Levy, Emily McTernan, Polly Mitchell, Erin Nash, Tristram Oliver-Skuse, Gavin Phillipson, Ben Steyn, as well as audiences at Cal State Fullerton, The University of Warwick, University College London, the University of Genoa, and a 'Mancept' Panel at the University of Manchester. Thanks, respectively, to Josh DiPaolo, Andrew Cooper, Jeff Howard, Corrado Fumagalli, and Maxime Lepoutre and Jonathan Seglow, for these invitations. I also had a helpful discussion about the ideas in this article on an episode of Suzanne Whitten's 'Fire in a Crowded Theatre' podcast.

References

Alexander, Gerard. 2006. Hear no evil, speak no evil. *CBS News,* 6th April 2006. cbsnews.com/news/hear-no-evil-speak-no-evil

Appiah, Kwame Anthony. 2012. What's wrong with defamation of religion? Pp. 164-182 in *The Content & Context of Hate Speech*, ed. Michael Herz and Peter Molnar. Cambridge: Cambridge University Press.

Ayers, Jessica D.; Diego Guevara Beltrán; Andrew van Horn; Lee Cronk' Hector Hurmuz-Sklias; Peter Todd; and Athena Aktipis. 2023. COVID-19 and friendships: agreeableness and neuroticism are associated with more concern about COVID-19 and friends' risky behaviors. *Personality and Individual Differences,* 213. https://doi.org/10.1016/j.paid.2023.112297

Bail, Chris. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton, NJ: Princeton University Press.

Barendt, Eric; Laurence Lustgarten; Kenneth Norrie; and Hugh Stephenson. 1997. *Libel and the Media: the Chilling Effect*. Oxford: Clarendon Press.

Bedi, Suneal. 2021. The myth of the chilling effect. *Harvard Journal of Law and Technology*, 35(1): 267-307. https://jolt.law.harvard.edu/assets/articlePDFs/v35/Bedi-The-Myth-of-the-Chilling-Effect.pdf

Bejan, Teresa M. 2017. *Mere Civility*. Cambridge, MA: Harvard University Press.

Blackford, Russell. 2019. *The Tyranny of Opinion: Conformity and the Future of Liberalism*. London: Bloomsbury.

Bor, Alexander and Michael Bang Peterson. 2022. The psychology of online political hostility: a comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, 116(1): 1-18. https://doi.org/10.1017/S0003055421000885

Brink, David O. 2001. Millian principles, freedom of expression, and hate speech. *Legal Theory*, 7(2): 119-157. https://doi.org/10.1017/S1352325201072019

Brown, Alexander. 2015. *Hate Speech Law: A Philosophical Examination*. New York: Routledge.

Cesarini, David; Christopher T. Dawes; Magnus Johannesson; Paul Lichtenstein; and Björn Wallace. 2009. Genetic variation in preferences for giving and risk taking. *Quarterly Journal of Economics,* 124(2): 809-842. https://doi.org/10.1162/qjec.2009.124.2.809

Cohen, Joshua. 1993. Freedom of expression. *Philosophy & Public Affairs*, 22(3): 207-263. https://www.jstor.org/stable/2265305

Connolly, P. J. 2021. Trolling as a speech act. *Journal of Social Philosophy,* 53(5): 404-420. https://doi.org/10.1111/josp.12427

Deckers, Jan and Jonathan Coulter. 2022. What is wrong with the international holocaust remembrance alliance's definition of antisemitism? *Res Publica,* 28(4): 733-752. https://doi.org/10.1007/s11158-022-09553-4

Doris, John. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Festenstein, Matthew. 2018. Self-censorship for democrats. *European Journal of Political Theory*, 17(3): 324-342. https://doi.org/10.1177/1474885115587480

Fuller, Lon L. 1958. Positivism and fidelity to law – a reply to Professor Hart. *Harvard Law Review*, 71(4): 630-672. https://doi.org/10.2307/1338226

Gaus, Gerald. 2010. *The Order of Public Reason*. Cambridge: Cambridge University Press.

Gould, Rebecca Ruth. 2022. Legal form and legal legitimacy: the IHRA definition of antisemitism as a case study in censored speech. *Law, Culture and the Humanities*, 18(1): 153-186. https://doi.org/10.1177/174387211878066

Gray, John. 1991. Introduction. Pp. vii-xxx in *John Stuart Mill: On Liberty and Other Essays*, ed. John Gray. Oxford: Oxford University Press.

Harman, Gilbert. 1999. Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, 99: 315-331. https://doi.org/10.1111/1467-9264.00062

Hart, H. L. A. 1958. Positivism and the separation of law and morals. *Harvard Law Review*, 71(4): 593-629. https://doi.org/10.2307/1338225

Hazlett, Thomas and David Sosa. 1997. Was the fairness doctrine a chilling effect? Evidence from the postderegulation radio market. *Journal of Legal Studies*, 26(1): 279-301. https://www.jstor.org/stable/10.1086/467996

Heinze, Eric. 2016. *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.

Hemel, Daniel and Ariel Porat. 2019. Free speech and cheap talk. *Journal of Legal Analysis*, 11(1): 46-103. https://doi.org/10.1093/jla/laz004

Hong, Ryan and Sampo Paunonen. 2009. Personality traits and health-risk behaviours in university students. *European Journal of Personality*, 23(8): 675-696. https://doi.org/10.1002/per.736

Horton, John. 2011. Self-censorship. *Res Publica*, 17(1): 91-106. https://link.springer.com/article/10.1007/s11158-011-9145-3

Jacobs, Laura and Joost van Spanje. 2020. Prosecuted, yet popular? Hate speech prosecution of anti-immigration politicians in the news and electoral support. *Comparative European Politics*, 18: 899-924. https://doi.org/10.1057/s41295-020-00215-4

Morgenroth, Thekla; Michelle K. Ryan; and Cordelia Fine. 2022. The gendered consequences of risk-taking at work: are women averse to risk or to poor consequences? *Psychology of Women Quarterly*, 46(3): 257-277. https://doi.org/10.1177/03616843221084048

Jensen-Campbell, Lauri A. and William G. Graziano. 2001. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality*, 69(2): 323-361. https://doi.org/10.1111/1467-6494.00148

Jensen-Campbell, Lauri A.; Katie A. Gleason; Ryan Adams; and Kenya T. Malcolm. 2003. Interpersonal conflict, agreeableness, and personality development. *Journal of Personality*, 71(6): 1059-1086. https://doi.org/10.1111/1467-6494.7106007

Jones, Mariette. 2019. The Defamation Act 2013: a free speech retrospective. *Communications Law*, 24(3): 117-131. https://api.semanticscholar.org/CorpusID:214186531

Joseph, Elizabeth D. and Don Zhang. 2021. Personality profile of risk-takers: an examination of the big five facets. *Journal of Individual Differences*, 42(4): 194-203. https://doi.org/10.1027/1614-0001/a000346

Kahneman, Daniel and Amos Tversky. 1979. Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2): 263-292. http://dx.doi.org/10.2307/1914185

Kendrick, Leslie. 2013. Speech, intent, and the chilling effect. *William & Mary Law Review*, 54(5): 1633-1691. https://scholarship.law.wm.edu/wmlr/vol54/iss5/4

Kenyon, Andrew. 2006. *Defamation: Comparative Law and Practice*. Abingdon: UCL Press.

Malik, Nesrine. 2019. The myth of the free speech crisis. *The Guardian*, 3rd September

2019. https://www.theguardian.com/world/2019/sep/03/the-myth-of-the-free-speech-crisis

Mason, Lilliana. 2018. *Uncivil Agreement: How Politics Became Our Identity*. Chicago: University of Chicago Press.

Meiklejohn, Alexander. 1948. *Free Speech and its Relation to Self-Government*. New York: Harper & Brothers.

Nelson, Cary. 2021. Accommodating the new antisemitism: a critique of 'the Jerusalem declaration'. *Fathom*, April 2021. https://www.fathomjournal.org/fathom-long-read-accommodating-the-new-antisemitism-a-critique-of-the-jerusalem-declaration

Nicholson, Nigel; Emma Soane; Mark Fenton-O'Creevy; and Paul Willman. 2005. Personality and domain-specific risk taking. *Journal of Risk Research*, 8(2): 157-176. https://doi.org/10.1080/1366987032000123856

Norris, Pippa. 2023. Cancel culture: myth or reality? *Political Studies*, 71(1): 145-174. https://doi.org/10.1177/00323217211037023

Olberding, Amy. 2019. *The Wrong of Rudeness: Learning Modern Civility from Ancient Chinese Philosophy*. Oxford: Oxford University Press.

Penney, Jonathon W. 2022. Understanding chilling effects. *Minnesota Law Review*. 106(3): 1451-1530. https://minnesotalawreview.org/wp-content/uploads/2022/04/6-Penney_Web.pdf

Quong, Jonathan. 2011. *Liberalism without Perfection*. Oxford: Oxford University Press.

Raz, Joseph. 1988. *The Morality of Freedom*. Oxford: Clarendon Press.

Rill, Leslie; Elizabeth Baiocchi; Megan Hopper; Katherine Denker; and Loreen N. Olson. 2009. Exploration of the relationship between self-esteem, commitment, and verbal aggressiveness in romantic dating relationships. *Communication Reports*, 22(2): 102-113. https://doi.org/10.1080/08934210903061587

Russell, Bertrand. 1998. The triumph of stupidity. Pp. 27-28 in *Mortals and Others, Volume II: American Essays 1931-1935*, ed. Harry Ruja. New York: Routledge.

Salameh, Anas A.; Hameeda Akhtar; Rani Gul; Abdullah Bin Omar; and Sobia Hanif. 2022. Personality traits and entrepreneurial intentions: financial risk-taking as mediator. *Frontiers in Psychology*, 13, article 922718. https://doi.org/10.3389/fpsyg.2022.927718.

Saunders, Joe. 2019. Lynton Crosby and the dark arts of democracy. Pp. 53-68 in *Media Ethics, Free Speech, and the Requirements of Democracy*, ed. Carl Fox and Joe Saunders. New York: Routledge.

Schauer, Frederick. 1978. Fear, risk, and the first amendment: unraveling the 'chilling effect'. *Boston University Law Review*, 58(5): 685-732. https://heinonline.org/HOL/P?h=hein.journals/bulr58&i=695

_____ 2020. The hostile audience revisited. Pp. 65-83 in *The Perilous Public Square: Structural Threats to Free Expression Today*, ed. David E. Pozen. New York: Columbia University Press.

Shaw, Kathryn L. 1996. An empirical analysis of risk aversion and income growth. *Journal of Labor Economics*, 14(4): 626-653. https://www.jstor.org/stable/2535442

Shiffrin, Steven H. 1978. Defamatory nonmedia speech and first amendment methodology. *UCLA Law Review,* 25(5): 915-963. https://scholarship.law.cornell.edu/facpub/1174

Simpson, Robert Mark. 2018. Regulating offense, nurturing offense. *Politics, Philosophy, and Economics*, 17(3): 235-256. https://doi.org/10.1177/1470594X17741228

Sims, Ceri M. 2017. Do the big-five personality traits predict empathic listening and assertive communication? *International Journal of Listening*, 31(3): 163-88. https://doi.org/10.1080/10904018.2016.1202770

Soane, Emma and Nik Chmiel. 2005. Are risk preferences consistent? The influence of decision domain and personality. *Personality and Individual Differences*, 38(8): 1781-1791. https://doi.org/10.1016/j.paid.2004.10.005

Srinivasan, Amia. 2018. The aptness of anger. *Journal of Political Philosophy*, 26(2): 123-44. https://doi.org/10.1111/jopp.12130

Stern, Kenneth. 2019. I drafted the definition of antisemitism; rightwing Jews are weaponizing it. *The Guardian*, 13th December 2019, https://www.theguardian.com/commentisfree/2019/dec/13/antisemitism-executive-order-trump-chilling-effect

Strossen, Nadine. 2018. *Hate: Why We Should Resist it With Free Speech, Not Censorship*. Oxford: Oxford University Press.

Sunstein, Cass. 2020. Falsehoods and the First Amendment. *Harvard Journal of Law and Technology*, 33(2): 387-426. https://jolt.law.harvard.edu/assets/articlePDFs/v33/33HarvJLTech387.pdf

Townend, Judith. 2017. Freedom of expression and the chilling effect. Pp. 73-82 in *The Routledge Companion to Media and Human Rights*, ed. Howard Tumber and Silvio Waisbord. London: Routledge.

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

Williams, Daniel. 2023. The marketplace of rationalizations. *Economics and Philosophy*, 39(1): 99-123. https://doi.org/10.1017/S0266267121000389

Wolfson, Nicholas. 1997. *Hate Speech, Sex Speech, Free Speech*. Westport, CT: Praeger.

Youn, Monica. 2013. The chilling effect and the problem of private action. *Vanderbilt Law Review*, 66(5): 1473-1540. https://scholarship.law.vanderbilt.edu/vlr/vol66/iss5/3