Philosophy & Public Affairs

ORIGINAL ARTICLE OPEN ACCESS

What's Wrong With Self-Censorship?

Gideon Elford

Department of Politics and International Relations, University of Oxford, Oxford, England

Correspondence: Gideon Elford (gideon.elford@new.ox.ac.uk)

Received: 4 July 2025 | **Revised:** 24 May 2026 | **Accepted:** 25 May 2026

Keywords: authenticity | freedom of speech | moral autonomy | self-censorship

ABSTRACT

In recent years, discourse on freedom of speech has shifted away from exclusive focus on the state and towards societal threats to speech. Amidst this change, the notion of "self-censorship" has gained increased prominence. Not only has self-censorship emerged as a common reference point, several recent studies identify it as embodying a corrosion of speech-related values. Despite this, the term self-censorship lacks a clear and settled definition, and the normative status of self-censorship isn't self-evident. While it can be troubling if people feel unable to speak openly, judicious silence is often a virtue, and sometimes morally required. Against this backdrop, I offer an account of a central bad-making dimension of self-censorship that helps us distinguish benign from pernicious forms. On this view, self-censorship is pernicious to the extent that it involves acquiescing to others' judgments of acceptable speech or belief. Identifying what is pernicious about certain forms of self-censorship in turn helps us sort between the types of social penalties on speech and belief that we ought to be more or less concerned about.

1 | Introduction

The problem with freedom of speech is that it affords people the license to say horrible things. Indeed, one could be forgiven for viewing much contemporary debate about free speech as disagreement over how far the manifest harms of speech should be weighed against disadvantages of regulating it. Everyone agrees that people can say and do the terrible with their words, it is just that the drawbacks of censorship are also pretty profound. Understood this way, one's view of speech regulation is a reflection of which of these is seen as the lesser of two evils. Framed in these terms, it seems like something like *self*-censorship presents a fruitful potential solution. After all, if directed appropriately, self-censorship promises all of the good—restraint on harmful speech—without the autonomy-damaging drawbacks of censorship *simpliciteur*. No doubt the operative "if" is a big one, but on its face self-censorship appears a force for good. Given that freedom of speech entails the freedom to remain silent, self-censorship might profitably marry free-speech-in-action with a self-discipline that restrains expressive harm.

And yet, self-censorship is more often spotlighted as a villainous character—a shadowy accomplice to the more overtly censorious threats to speech. Often in relation to social sanctions, self-censorship has emerged as a common reference point, with several recent studies labelling it a salient and growing threat to speech [1–8]. The real damage wrought by the infliction of social penalties on unpopular speech, it is said, is not felt by those who suffer the punishment themselves but exacted through fear instilled in those for whom such consequences serve as a warning; spreading reluctance to speak, in dread of inviting sanction on themselves. So which is it, is self-censorship valuable or lamentable? It seems troubling if people feel unable to speak openly, yet judicious silence is often a virtue, and in many cases morally required. Social sanctions that discourage speech can appear oppressive but social life naturally, and reasonably, involves facing some social costs for what we say and do. Self-censorship is routinely presented as a free speech concern, yet the very term convicts the "self" as the culprit exacting the censorship. It isn't self-evident whose speech is really unfree. Even so, there does seem to be something morally troubling about self-censorship.

But how can we make sense of its redeeming potential alongside its understandably murky reputation?

Despite the fact that the notion of self-censorship has occupied a central place in our vocabulary regarding freedom of speech, and embodies this intriguing ambivalence, it has enjoyed little extended philosophical treatment.[1] In particular, scarcely anything has been written to help us distinguish benign forms of self-censorship from more pernicious cases. Here I offer an account that offers just such a guide, under which, roughly, self-censorship is pernicious to the extent that it involves persons acquiescing to others' views concerning acceptable speech or belief. I develop this account as follows. In Section 2, I specify the concept of self-censorship that I take for granted, and note an important feature necessary to distinguish it from censorship in general. In Section 3, I detail some existing philosophical work on self-censorship and contend that we presently lack a compelling account that distinguishes benign from pernicious forms of self-censorship. In Section 4, I offer my account of self-censorship as acquiescence and contend that it offers an intuitively compelling basis for sorting pernicious from benign forms. In Section 5, I contend that this form of self-censorship is pernicious because it involves abdicating moral autonomy over speech and, to that extent, it involves resigning expressive independence. In Section 6, I connect this analysis to the conditions liable to encourage self-censorship and argue that it helps us understand which kinds of social sanction are likely to be justified. In turn, I contend that we have moral reasons to refrain from foreseeably and avoidably contributing to a discursive culture that fosters acquiescent self-censorship. In Section 7, I address some objections and Section 8 concludes.

2 | Self-Censorship

The concept of self-censorship is ripe for conceptual interrogation, particularly given the array of meanings to which people affix the label.[2] That isn't my primary task here, however. One very broad rendering of "self-censorship" takes it to cover all cases in which a person intentionally constrains their own speech. I offer a slightly narrower rendering of the concept, however. For censorship to be distinctively *self*-censorship, a meaningful role for the self is required [11], p. 47. The maiden who fears a fiery death at the stake for confessing her faith does not so much *censor herself*, rather she shrinks from the censorship of others. A contrast with censorship *per se* is interesting precisely because, for *self*-censorship, the self is not entirely off-stage but plays a substantial role in the restraint of speech [12], pp. 98, 99. Crucially, the person enjoys reasonable alternatives to remaining silent. Self-censorship, then, involves persons *voluntarily* constraining their own speech.[3]

Though a comprehensive account of conditions for voluntariness escapes adequate defense here, some further remarks are in order. A person refrains from speech non-voluntarily to the extent that speech entails (risk of) unreasonable consequences.[4] Clearly, the reasonableness of those consequences depends substantially on the gravity of costs they entail. Death at the stake is a cost of vastly greater significance than, say, a mild rebuke from friends. But consequences can be unreasonable for reasons other than narrowly self-affecting costs. In an oppressive dictatorship, I might refrain from criticizing the ruling powers by dint of risks to friends and family, judged guilty by association, rather than jeopardy for my own safety specifically.

Moreover, what matters for voluntariness is not strictly whether the consequences of speech are unreasonable but whether it is reasonable to judge that they are. If a gun-wielding person demands my silence, then I keep quiet non-voluntarily, even if it transpires the gun was merely a convincing replica. Even though the consequences for speaking were not in fact unreasonable, it was reasonable to judge them such. Similarly, if I reasonably judge that it will destroy my career to express a certain view, my silence is non-voluntary even if I enjoy safeguards against suffering that fate that I couldn't have reasonably anticipated.[5]

Beyond this sketch of a distinction between voluntary and non-voluntary silence, there are two further salient considerations that I refrain from adopting a view on here. First, the degree of cost that renders speech unreasonable. As I intimate, there are clear cases on either side—death and career destruction are unreasonable, mild rebukes reasonable. Where to draw the line in between those cases is a question I leave open. Second, whether *unreasonable* mistaken judgments about consequences render silence non-voluntary. Reasonable but mistaken fears of career destruction plausibly undercut voluntariness of speech, but it is more contestable whether groundless and irrational fears do so. Either way, my account is compatible with disagreement over either issue.

If these gaps diminish the comprehensiveness of the account, they nevertheless have something to commend them. One reason that the voluntary character of self-censorship is significant is that it convicts individuals themselves of responsibility for restraining their speech. The difference between career destruction and mild rebuke cases is that in the latter I enjoy space to exercise autonomous discretion over my silence, arbitrating between competing considerations that can each lay claim to my attention over the other. It is my "self" that censors by playing a formative role in determining whether I should speak. In consequence, responsibility for not speaking is partly, but inescapably, my own [12], p. 102. By leaving unspecified the precise conditions for voluntariness, the account of self-censorship I offer below is compatible with different views of the preconditions for responsible choice. And it is the voluntary dimension of self-censorship that makes any would-be negative character more challenging to explain. The external suppression of speech can be readily analyzed in terms of the gamut of reasons proffered for why *freedom* of speech is of value. When the agent is responsible for restraining their own speech, however, any badness or wrongfulness demands further explanation.

My limited conceptual prescription, then, is that *self*-censorship is voluntary. That prescription does not, of course, entail that all non-voluntary suppression of speech is properly classified as "censorship." The criteria for censorship proper is another matter I leave open. Still, it is instructive to briefly note an alternative way of dividing self-censorship from censorship proper. This is based on the idea that censorship does not merely suppress speech but disallows it.[6] Censorship might be distinguished from self-censorship on the grounds that the former involves an exercise of external authority. Take state censorship,

for instance. State censorship does not merely punish speech but *prohibits* it [26]. By censoring, the state claims authority to determine the bounds of permissible speech. So, the supposition goes, self-censorship includes those cases where external authoritative determination of permissible speech is absent. Those are the cases in which individuals themselves retain authoritative discretion over what they should say.

Though my account does not make the cut between censorship and self-censorship on this basis, the notion of authority is central to the account. Self-censorship, I will argue, is especially regrettable when it entails acceding to standards of acceptable speech or belief imposed by others' judgments. It involves effectively surrendering to others' determination of permissible speech, not by affirming their authority to do so but by yielding to it. Like citizens of an occupied territory complying with but never truly obeying the alien rule, this form of self-censorship involves a person reluctantly submitting to standards of appropriate speech set by others' judgments that, if only for more courage and fortitude, she might have steeled herself to resist. In the coming section I more fully explain that bad-making feature of self-censorship and contend that it successfully distinguishes benign from pernicious cases.

3 | Benign and Pernicious Self-Censorship

We routinely exact editorial control over our speech—to communicate clearly, to obviate misunderstanding, to provoke emotional effect. Trimming, cutting, and shaping our expression involves intentionally holding back from using certain words and not others. Such is an integral part of writing of almost all forms. There is nothing inherently troubling about this; indeed, some will doubt whether anything of this sort counts as self-censorship at all. We also curate our language for moral reasons. This is more readily classified as self-censorship, although it remains untroubling as a general matter. However, while much of the control we exercise over our own speech is not of concern, other cases are worrisome. But what distinguishes troubling from untroubling cases?

The question has received only limited philosophical attention. Existing work offers some insight but extant views are either limited in detail or else face certain drawbacks as explanations of the perniciousness of self-censorship. John Horton offers a careful conceptual analysis of self-censorship and rightly observes that the cases of self-censorship that concern us are those where the self is both the author and the instrument of censorship [12]. As I remarked above, in order to be self-censorship, rather than censorship proper, there must be a sense in which the person voluntarily restrains their speech. Yet even voluntary expressive restraint seems disquieting when it results from some degree of external pressure. This tension between an external coercive element and internal self-restraint is, as Horton puts it, "at the heart of the practice of self-censorship—a form of action in which self-determination and heteronomy co-exist in an uneasy but inextricable relationship." [12], p. 105. That tension partly explains why, as Horton aptly remarks, self-censorship is generally thought to be "under some sort of moral cloud." [12], p. 97. This account tells us that self-censorship occupies that expressive territory within which there is pressure against speaking in certain ways but which does not rise to the level of constituting irresistible pressure against doing so. Self-censorship in such cases is morally troubling but not self-evidently so. Beyond this insight, however, Horton explicitly refrains from pursuing a substantive analysis to help us identify pernicious instances of self-censorship.

Matthew Festenstein offers a promising advance on Horton's view. He contends that self-censorship is contemptible when refraining from speech is "underpinned by a certain kind of explanation, in terms of a problematic power or influence relationship." [10], p. 326. On this view, we differentiate between benign and pernicious cases of self-censorship by explaining the latter "in terms of the various agents' responses to the specific power relations in which they are enmeshed." [10], pp. 326, 327. Again, there is important wisdom in this account, but it has drawbacks as a guide to distinguishing concerning from innocuous self-censorship. First, it wholly depends on a broader account of which power relations are problematic. This is not necessarily a weakness of the account as such. After all, it offers us a broad framework within which disagreements about the normative character of self-censorship can be housed. Indeed, Festenstein sometimes presents his view in those terms—as a way of framing contested intuitions about cases of self-censorship [10], p. 329. Our disagreement over the nature of power and its normative character is naturally reflected in our judgments of which cases of self-censorship are troubling. As a result, though, Festenstein's account takes us not much further than pointing to the type of reason that would make self-censorship pernicious, rather than delivering a substantive account of the particular reasons.

Second, the breadth of the framework flattens crucial differences between distinct dynamics by which persons refrain from speech. Festenstein adopts a capacious view of the variety of ways that power relations can suppress speech, encompassing, "coercion, ideological domination, adaptive preference formation, incentives, influence, and so on." [10], p. 327. The problem with this is that power operates to suppress speech in markedly different ways, and the critical explanation for silence in different cases may be troubling for very different reasons.[7]

Consider a pair of cases. Deborah is brow-beaten with the message that atheism is "evil heresy" so, despite holding atheist beliefs, she refrains from voicing them lest she meet with uncomfortable peer disapproval. Hercule faces the same anti-atheist message but comes to internalize that same view of atheism, not through rational persuasion, but by having his sensibilities cudgeled into line. On Festenstein's account, Hercule's reconfigured sense of appropriate speech counts as contemptible self-censorship. His conditioning plausibly reflects a problematic power relation that explains why he no longer expresses pro-atheist sentiments. And yet, even if this is a case of self-censorship, the reasons to regret Deborah's self-censorship seem importantly different. Whereas Hector doesn't espouse atheism because he now takes a different view about what he should say, Deborah is discouraged from speech by the specter of social consequences. The dynamic by which speech is suppressed in Deborah's case involves self-constraint, whereas, for Hercule, problematic power produces silence by transforming his perspective.

Of course, we might refine Festenstein's account to narrow the range of ways in which power operates to limit speech. Perhaps Hercule's case feels substantially different from Deborah's because it lacks the requisite tension for self-censorship—now he is a true believer; there is no tug between what Hercule intends to say and the censorship he imposes on himself. Yet even on a narrower account—with tug in play—there are still important differences in how power operates to stifle expression. Consider, for instance, Harriet, who is so repeatedly told of her ignorance and stupidity that she lacks confidence that she has anything worthwhile to say. She sincerely believes that she shouldn't speak because what she is otherwise inclined to say simply lacks value. Again, this seems like a case of contemptible self-censorship on Festenstein's account, but it looks importantly different from a case where, for instance, Harriet refrains from saying things she believes have value but fears being regarded as stupid by those she might say them to. These limitations of the Festenstein account are understandable, given that it offers a very general criterion by which to distinguish bad from innocuous self-censorship. It does indicate, however, that it is at most a staging post *en route* to a more thorough excavation.

The third issue with the Festenstein account, though, is a result of its externalist character. That is, its distinction between pernicious and benign self-censorship depends heavily on the character of surrounding power relations. It emphasizes the conditions that give rise to self-censorship, rather than the character of the self-censorship itself. This makes it difficult to account for perniciousness in cases where individuals self-censor when they mistakenly, but reasonably, fear reprisal or consequences for speaking up. Suppose Nathan has good reason to believe the stories percolating his College campus that professors will grade you down for voicing certain political opinions and, fearing this, keeps his views to himself. In fact, no such suspect power relation, or likely exercise, in fact exists. Still, our reasons for being concerned about Nathan's self-censorship are not entirely extinguished by the fact that problematic external relations do not in fact obtain.

Moreover, when people complain about self-censorship they are usually not simply saying that morally suspect social conditions suppress speech. To be sure, when self-censorship is decried as a salient social ill, critics often target the oppressive dimension of a culture that fosters it. Still, they are also contending that its propensity to foster self-censorship is itself a reason for regarding that culture as oppressive. There is something pernicious about self-censorship that *explains* the badness of the social relations, rather than merely reflecting it. So, even if external problematic power relations are a pernicious-making source of certain forms of self-censorship, it remains important to explore the internal reasons persons have for silence.

To understand pernicious cases of self-censorship we must also attend to its character, rather than exclusively on the conditions that give rise to it. This is also the approach taken by J. P. Messina, who distinguishes "self-censorship" from "self-restraint" based on the reasons for which persons withhold their speech [13], p. 31. For Messina, "[w]hereas self-censorship is grounded in fears of bearing personal costs for expression … *mere self-restraint* is based on authentic judgment that what one wants to express is for some reason inapt." [13], pp. 30–31, original emphasis. Though Messina intends this primarily as a conceptual move, it could be understood as a distinction between pernicious and benign forms of self-censorship.

Read in those terms, Messina's view captures something important—that one's judgments of the aptness of speech bear significantly on the perniciousness of self-censorship. However, it remains a very general gloss on the distinction and does not itself explain why the distinction should be drawn in roughly this place. Moreover, as I will explain, some benign cases of self-censorship involve fears of personal costs whereas others are pernicious even though the feared costs are not personal ones. My account develops and refines the core insight of Messina's distinction and explains when and why an agent's own judgments concerning the aptness of speech bear on the perniciousness of self-censorship.

4 | Self-Censorship as Acquiescence

In headline form, I argue that self-censorship involves a distinctive form of perniciousness to the extent that it involves acquiescing to others' judgments concerning the views one ought to hold or express. Acquiescence entails succumbing to pressure to conform with others' judgment about what to say or believe in spite of the fact that one does not in fact agree with that judgment or one does not accept the authority of their status as judge. A person self-censors in this sense when they refrain from speech in order to avoid anticipated negative consequences constituted by or resulting from others' negative judgments while also not sharing those judgments, nor judging them fitting reasons to refrain from speech or belief. In this section I further elaborate this view of self-censorship as acquiescence, argue that it allows us to distinguish benign from pernicious cases, and explain what makes this form of self-censorship lamentable. In the section that follows I defend the advantages of this acquiescence view over certain rival explanations.

This account deals with clearly troubling and untroubling cases as you would hope; sorting them into appropriate categories. Being polite in order to spare someone's feelings is a benign case on this view. Consider:

Dinner Party—Nate, a fierce opponent of state gun control, is at a dinner party with friends. In the course of conversation, it transpires that another guest, Matilda, recently lost her child in a mass shooting. When the conversation turns to the events in question, a clearly distressed Matilda voices an impassioned case in favor of greater gun regulation. In other circumstances Nate may have protested, but he judges it improper to make a case against Matilda's view. Out of respect for her still raw trauma, he remains quiet.

While Nate keeps silent in order to avoid a negative consequence—Matilda's distress—he nevertheless restrains his own speech for what he takes to be sound moral reasons. Even though the negative consequences issue from Matilda's judgments that he doesn't share—the view of gun regulation—this isn't a case of Nate allowing Matilda's beliefs to govern the permissibility of his own speech. Instead, he affirms a morally authoritative principle concerning when to speak that accounts

for her beliefs as a morally relevant consideration falling under it. He doesn't fear her judgment and the ramifications of outwardly contradicting it but accords it moral weight in his decision not to speak.

Compare a paradigmatic case of pernicious self-censorship:

Hostile Dinner Party—Nate is at another dinner party with an entirely different group of friends and the conversation turns to gun control. The only people who speak are vociferous in support of gun control, labeling opponents "irrational," "callous," and "ideological radicals." The dominant tone of the conversation is a palpable contempt for anyone opposed to gun control, to a point where Nate reasonably speculates that he would be told to leave if his views became known.

Nate doesn't reveal his position on gun control. He refrains from speaking specifically to avoid the contempt and hostility that airing his views would, he judges, likely incur. Here others' judgments concerning acceptable beliefs set the terms of discourse. Unlike *Dinner Party*, others' beliefs do not function as moral considerations that Nate takes account of but take the form of potential sanctions he resolves to avoid. As I will explain in what follows: what makes it distinctively morally troubling *qua* self-censorship is that Nate refrains from speech in order to avoid subjection to others' judgment that he does not share.

Dinner Party and *Hostile Dinner Party* represent clear cases of benign and pernicious self-censorship respectively. A crucial test of the account, though, is whether it can diagnose less intuitively clear cases. Suppose *Hostile Dinner Party* is a *Birthday Celebration* for Micah, one of Nate's close friends, who happens to be apolitical. Now suppose that Nate's silence is motivated primarily by a concern to avoid ruining Micah's birthday dinner—that challenging the pro-gun control consensus would certainly do. Nate refrains from speaking for moral reasons, but those moral reasons pertain to the very same beliefs of others that create social pressure. Is this benign verbal restraint or worrying kowtowing to others' judgments?

For good reason, *Birthday Celebration* is less intuitively clear. As it stands, though, Messina's cut—between fearing personal costs (pernicious) and viewing the speech inapt (benign) – offers only limited guidance here. On one hand, Nate fears the consequences of others' judgments for the dinner party; on the other hand, he affirms Micah's potential distress as a fitting moral reason to refrain from speaking. So both fear of costs and inaptness are in play. I contend that despite moral considerations informing Nate's resolution to remain silent, there is still a pernicious dimension to his self-censorship and we can understand how those moral considerations do not wholly launder his reasons for self-censorship by attending to how others' judgments figure in Nate's decision not to speak. Micah's feelings are an operative moral reason for Nate only because they are brought into play by other guests' judgments whose disruptive potential he fears inviting. Nate judges Micah's feelings a good moral reason to remain silent, but this is different from Nate judging his own silence as a fitting consideration necessary to protect Micah's feelings. Rather, (let us imagine) Nate judges it illegitimate if guests' antagonistic reaction to his speech were to disrupt Micah's party.

Further tweaking the case to *Birthday Celebration 2* supports this verdict. Suppose Nate doesn't anticipate a hostile reaction from the pro-gun control crowd, but he does suspect that some will find his views distasteful and upsetting. In other circumstances he would nevertheless present his beliefs, but he is mindful that Micah is desperate for all his guests to enjoy their time, and the guests' preferences for frictionless dinner conversation are reasonable. For Micah's sake, he keeps quiet. This now more closely resembles untroubling principled silence rather than pernicious self-censorship. Nate is not only motivated by a consideration he takes to be morally relevant (Micah's feelings); that consideration is operative because of considerations he also sees as reasonable (guests' enjoyment). The pivotal difference between *Birthday Celebration* and *Birthday Celebration 2*, then, is how Nate apprehends the role of others' judgments in recommending his silence. In both cases, Nate keeps silent for moral reasons—concern for Micah's feelings—but in the latter, he apprehends the responses of others, and the judgments they are based on, as reasonable determinants of Micah's distress. By contrast, in the former, he comprehends their anticipated responses to his speech as hostile forces to be assuaged and placated. In this case, Nate allows others' judgments to set the appropriate terms of discourse in a way he judges unreasonable but gives way to nonetheless.

Other cases are more straightforwardly benign, and my account is able to make good sense of why they do not trouble us in the way that other cases do.

Bargaining—Garrett doesn't reveal his true opinion of the used car that he's selling, because he worries that Alyssa won't buy it from him in that case. He restrains his speech in order to avoid potential personal cost, but it isn't a cost inflicted because Alyssa judges his speech or belief inappropriate, and so it is rightly classified as benign self-censorship on my account.

Confession—Sally witnesses a crime and would ordinarily give evidence in support of a prosecution but refrains from doing so because it risks implicating herself in criminal malfeasance that she was in fact party to. She keeps quiet for fear of legal ramifications.

Interview—Mariane is asked during an interview about her most serious past work-related mistake. She knows that the honest answer involves a revelation that will substantially reduce her chance of successfully getting the job. This is because the mistake involved a basic error in judgment that calls into question her fitness for the job. Instead, she proffers a less damaging anecdote.

Each case involves an individual voluntarily withholding their own speech in order to avoid personal cost. And yet, none seem especially pernicious as cases of self-censorship. Again, Messina's distinction (if read as a cut between pernicious and benign cases) isn't drawn in quite the right place. Mere prudential fear of the consequences of one's speech isn't enough to render silence troubling. Missing is the further element—that the consequences are issued by others' judgments as to what the person ought to say or believe, despite those judgments not being shared by the self-censoring person, nor deemed fitting reasons to refrain from speech.

To draw this out, consider:

Ideological Interview—Mariane is asked the same interview question about past work-related mistakes. Again, she knows that her chances are greatly reduced if she gives an honest answer, but in this case, it is because that answer reveals she previously worked for a socialist policy unit. She has reason to fear that the interviewer despises socialists. For this reason, she gives a different example.

My account explains why Mariane's self-censorship in *Ideological Interview* is troubling in contrast to her self-censorship in *Interview*. In both she curtails her speech to avoid falling out of consideration for the job. What makes the former pernicious, however, is that Mariane hides from a standard of acceptable belief that she does not share. A similar analysis can be given of *Confession*. Sally seeks to avoid consequences issuing from revelations concerning her culpability. However, those consequences are not rooted in others' judgments concerning acceptable speech or belief that she doesn't share. Suppose, instead, that Sally fears her confession will implicate her as an anti-abortion activist and worries that social condemnation will follow from those hostile to that position. Her self-censorship in that case appears distinctively pernicious in a way that avoidance of legal penalties for criminality is not.

Note, moreover, (pace Festenstein) that problematic power relations do not themselves transform benign self-censorship into pernicious. Suppose, in *Bargaining*, Garrett's exchange with Alyssa is an exploitative one where she is able to unjustly leverage her unfair bargaining power. Though the transaction may be unjust, it's not clear that Garrett's self-censorship is itself any worse because of that. Now, one might argue that things are different in the contrast between *Interview* and *Ideological Interview*. It could be argued that we can explain our concern for Mariane's self-censorship in the *Ideological Interview* in terms of the unjust power dynamic she is subjected to. That is, only in the latter case is Mariane vulnerable to political discrimination. However, I don't think this captures what is basically troubling about the case *qua* self-censorship. It can't be that in *Ideological Interview*, but not *Interview*, Mariane is *subjected to* political discrimination. After all, Mariane avoids actual political discrimination precisely by hiding her beliefs. Nor is it necessarily true that Mariane is only vulnerable to political discrimination in the former case—she could be vulnerable to political discrimination in the latter case, but this simply doesn't inform her self-censorship. So the claim must be that Mariane's vulnerability to political discrimination *informs* her self-censorship in *Ideological Interview* alone. The problem with this explanation, though, is that Mariane's self-censorship could be informed by the reasonable but mistaken belief that she will be discriminated against on the basis of her socialist beliefs. Even if the unjust power dynamic does not obtain, then, I contend that Mariane's self-censorship remains troubling. Though an externalist approach may rightly identify a troubling dimension of certain instances of self-censorship and thereby correctly sort some self-censorship cases as pernicious, it struggles to explain perniciousness in cases of reasonable mistakes.

Thus far I have argued that self-censorship is pernicious when the consequences we seek to avoid are issued by others' judgments about what we ought to say or believe that the self-censoring party does not share. In those cases, individuals acquiesce to standards of appropriate speech or belief set by others. Some crucial refinements to this remain, however. Even in cases where we don't concur with the judgments of others regarding what ought to be said or believed, self-censorship is non-pernicious when we either accept the authority of others to impose normative requirements concerning acceptable speech, or where we endorse second order reasons to accept the judgments of others regarding what to say or believe. This refinement accords with the intuitively benign character of a raft of cases.

Schoolteacher—Jeannette teaches a curriculum on evolutionary theory that is contrary to her creationist beliefs. Against her inclination to voice dissent in the classroom, she scrupulously adheres to the school syllabus and teaches classes affirming evolutionary theory. Though there are sanctions for deviating from the school curriculum, Jeannette restrains herself because she recognizes the school's authority to determine what the students learn.

The self-censorship in *Schoolteacher* is benign. Though the standards of acceptable speech are rooted in beliefs contrary to Jeannette's own, she affirms the normative authority of the speech rule to which she adheres. Compare:

Reluctant Teacher—Clarke is a primary school teacher. His school policy entails conveying a particular view of sexual morality to the students, a view deeply at odds with Clarke's view. When teaching, Clarke is strongly inclined to challenge the official school position, deeming it a profoundly inappropriate message to propagate to children. Fearing for his job, however, Clarke remains silent.

Clarke's self-censorship in *Reluctant Teacher* is crucially different from Jeannette's. Jeannette bases her censorship on a rule whose authority she accepts. She affirms second-order reasons to adhere to a speech rule and thereby autonomously affirms the limit on her speech. Clarke, in contrast, censors himself purely in order to avoid sanctions inflicted to enforce beliefs he rejects.

It is important to be clear about the specific sense in which Jeannette accepts the authority of the school to determine appropriate speech. She accepts the school's normative authority to impose requirements on acceptable speech. Normative authority matters precisely because it sets standards of acceptable speech that the self-censoring party themselves endorses.[8] Jeannette judges that she ought not contravene the curriculum requirements precisely because they are a manifestation of the school's (perceived) power to impose normative requirements of this kind. This is crucially different from accepting that the school merely has the right to enforce requirements on acceptable speech. It is quite consistent with pernicious self-censorship that the person self-censoring accepts that others have a right to inflict the consequences from which they cower. To recall *Hostile Dinner Party*, for instance, Nate might very well accept that fellow attendees have the right to voice their distaste for the views he holds, and that the host has a right to eject him. This does not mitigate the sense in which Nate's self-censorship is troubling. On the other hand, as I will further explain, accepting normative authority is crucial because it sustains a sense in

which the self-censoring party themselves sets the standards for acceptable speech.

Jeannette's self-censorship is benign because she endorses second order reasons to accept the school's normative authority, even though she rejects the substantive beliefs on which the standards are based. An alternative route to benign self-censorship involves endorsing second order reasons to affirm the substantive beliefs themselves. Consider:

Insecure Graduate—newly minted doctoral graduate Julius prepares for an upcoming conference talk. He tentatively holds a philosophical view that his former supervisor and longtime mentor tells him is philosophically incoherent. Accounting for their respective philosophical pedigree, Julius concludes his supervisor is likely correct, even though he remains unconvinced by her first-order philosophical reasons. He doesn't defend the view in his conference talk.

Julius defers to his mentor, but he doesn't acquiesce to her judgment. Rather, he takes himself to have good reason to hold her philosophical judgment to be authoritative (at least in comparison with his own). He exercises his judgment rather than resigning it.[9] Suppose, instead, Julius doesn't defer to his mentor's judgment, but he also knows that defending his own view at the conference will hurt her feelings. Now, if he refrains for that reason, he engages in benign principled silence—he takes her judgments to have a bearing on the morality of speaking. Consider, though:

Negative Reference—Julius doesn't give the talk because he fears his mentor will write him a negative reference.

On the one hand, this resembles pernicious acquiescence—her judgment inflicts consequences that, through silence, he aims to evade. Yet on the other hand, this case seems less troubling than *Hostile Dinner Party* and *Reluctant Teacher*. What makes Julius's self-censorship less pernicious than those cases isn't that he accepts his mentor's right to supply the reference she sees fit. That is, it's not simply that he recognizes the existence and value of the practice of academic reference provision. Rather, as I have drawn the case, we suspect that Julius respects the legitimacy of his mentor's judgment even though he disagrees with it. Though he does not endorse her judgment, he regards her reference as a sincere and considered judgment of his philosophical credentials. The specter of the bad reference looms as a negative consequence, but it garners a degree of recognition as a legitimate implication of his speech, albeit one Julius still hopes to avoid. This is brought out by:

Biased Reference—Aria plans to give a conference paper defending the philosophical respectability of conservatism. She is confident her paper offers a robust and plausible account; however, she knows her supervisor has notorious contempt for people with conservative views. Fearing that her supervisor's ideological contempt will be reflected in a negative reference, Aria decides against giving the paper.

I contend that Aria's self-censorship is more pernicious than Julius's. Both Julius and Aria recognize that the practice of reference writing is legitimate, and that the practice appropriately gives supervisors discretion to write the references they see fit. Aria, however, does not also accept the legitimacy of the considerations that inform the negative reference she hopes to avoid—namely contempt for conservative views.[10] But is Julius's self-censorship at all pernicious, or does his recognition of the legitimacy of his supervisor's judgment render his prudential silence entirely benign? The answer to this is important for guiding our general view of self-censorship, largely because self-censorship commonly entails persons keeping quiet to avoid the aggregative effect of individually morally permissible conduct.

Perhaps we can draw a lesson from *Birthday Celebration 2*. In contrast to *Birthday Celebration*, Nate's self-censorship was benign because he recognized guests' responses as reasonable determinants of Micah's distress. This might suggest that whenever the consequences a person resolves to avoid are issued by responses judged reasonable by that person, the self-censorship is wholly benign. I don't think this is quite right, however. Crucially, the reasonableness of others' responses was relevant because it had bearing on whether Nate judged Micah's potential distress a fitting moral reason to remain silent. It is less clear that merely prudential avoidance of reasonable responses from others renders the self-censorship entirely benign. Consider:

Friendships—Danika has firm religious beliefs but lives in a secular society where the vast majority of others reject her views. She fears that revealing her religious convictions will mean no one will want to befriend her. Even so, she accepts that persons conditioning their initiation of friendship on shared views about religion is eminently morally permissible.

Abortion—Clarissa believes that women ought to have the right to choose an abortion but, in her society, pro-choice advocacy is typically subject to robust criticism and condemnation. While Clarissa concedes that it is reasonable for others to express their sincere disagreement with her view, out of discomfort with the prospect of being a target of criticism, she opts to stay quiet.

I hazard that Danika and Clarissa's self-censorship remains troubling to some degree. However, these cases are also intuitively less concerning than *Hostile Dinner Party*. By the same token, the self-censorship would be more troubling had Danika feared coordinated ostracism from others, or Clarissa feared derogatory and hostile verbal attacks. Now, one analysis here is an externalist one—*Friendships* and *Abortion* are less troubling because the treatment that Danika and Clarissa are liable to receive is itself less pernicious. And, of course, if those fears of ostracism and attack reflect reality, then this is surely part of the badness of the state of affairs as a whole. A more compelling diagnosis of their self-censorship specifically, however, is that it is less pernicious in *Friendship* and *Abortion* precisely because the consequences they fear are ones they regard as having some degree of legitimate place as consequences of their speech. We feel more ambivalent about cases like *Friendships* and *Abortion* (and Julius's fear of his supervisor writing a sincere but unflattering reference), because although the parties still fear the judgments of others, they do not regard them as entirely misplaced. My account explains this ambivalence—it is because the acquiescence in such self-censorship is less full. We refrain from speech in order to avoid consequences issuing from others' beliefs that we do not share, but we also recognize that those beliefs, and the

consequences they issue, have some legitimate bearing on the case—not that they should inform what ought to be said (as in *Schoolteacher* where Jeannette grants normative authority, or *Insecure graduate* where Julius grants epistemic authority), but in the sense that we don't submit to terms of acceptable speech and belief that we regard as wholly unreasonable.

Let us corral the dimensions of self-censorship considered to this point. There are several considerations that bear on the relative pernicious or benign character of self-censorship.

1. The reasonably anticipated consequences of speech are based in others' judgments concerning the acceptability of belief or speech that one does not share.
2. The agent judges that they have first-order moral reasons to refrain from speech.
3. The agent judges that they have second-order reasons to accept the authority of the speech rule (not merely the authority to punish), or the authority of the judgment concerning the speech or belief.
4. The agent judges that the anticipated costs of speech are reasonable/unreasonable consequences.

1 is a pernicious-making feature of self-censorship. Where 1 is absent, self-censorship does not involve the distinctive form of perniciousness—as acquiescence—for which I argue. 2 and 3 are benign-making features, though how they interact with 4 determines the degree to which self-censorship is benign or pernicious overall. Similarly, how far 4 applies directly to 1, in the absence of other benign-making features, bears on the degree to which self-censorship is pernicious. We can represent this as follows:

The root of potential perniciousness is 1. This is the basis on which persons acquiesce to others' judgments concerning the acceptability of speech or belief. Where neither 2 nor 3 apply, self-censorship is to some degree pernicious; however, its perniciousness depends, under 4, on how far the self-censoring party regards the feared consequences as reasonable implications of their speech or belief. When regarded as reasonable (*Negative Reference, Friendships, Abortion*), the self-censorship is less pernicious than when regarded as unreasonable (*Hostile Dinner Party, Biased Reference*). Very roughly, perceived reasonableness is a benign-tilting aspect of self-censorship whereas perceived unreasonableness is pernicious-tilting.

Under 2, when an agent judges that they have first order moral reasons to refrain from speech this renders the self-censorship benign, unless those the moral considerations are introduced because of consequences, rooted in others' opposing views of speech or belief, that the agent judges unreasonable to impose. In that latter case, self-censorship still involves acquiescing to others' judgments concerning acceptable speech or belief (*Birthday Celebration*). So whether 2 embodies benign or pernicious self-censorship depends on 4—whether the moral considerations are implicated because of consequences deemed unreasonable.

Similarly, under 3, where an agent judges they have second order reasons to accept the authority of a speech rule or the authoritative judgment of others concerning the belief or speech, self-censorship is benign. If the agent self-censors because they accept the normative authority of the speech rule to which they adhere, then they do not acquiesce to others' judgments but affirm the appropriateness of the rule in spite of its being grounded in judgments that differ from their own (*Schoolteacher*). If the agent self-censors because they accept the epistemic authority of others (*Insecure graduate*), then, again, the self-censorship is rendered benign because the agent affirms the judgments (for second order reasons accepting epistemic authority) that counsel against speech.

But does perceived reasonableness have a bearing on perniciousness under 3 in the same way as it does under 2, in relation to first order moral reasons to self-censor? The answer is a qualified yes, but we need to be precise about how reasonableness figures in this context. Self-censorship involves perniciousness even under 3 if the self-censoring party does not regard the exercise of authority as reasonable. This is importantly different from regarding potential consequences for transgressing the authoritative standard of speech as unreasonable. The perceived reasonableness of potential *consequences* is not pertinent because, *ex hypothesi*, the agent does not keep quiet in order to avoid those consequences but because they affirm the authority of the speech rule.[11] A pair of examples will help.

Journalist—Marcus reports from the crime desk for a local newspaper. There is a newspaper policy not to publish stories that will damage community relations. Marcus has a story he is convinced will not have any negative effect on community relations but also strongly suspects that the notoriously risk-averse newspaper owner will judge that it does and likely remove Marcus from the crime desk if it is published.

Here Marcus affirms the authority of the speech rule he would suffer consequences for (being perceived to be) transgressing, but does not self-censor strictly to adhere to that speech rule but because he fears being penalized because of an unreasonable application of that rule. His self-censorship is pernicious to the extent that it involves acceding to judgments about the appropriateness of speech from which he dissents.

Journalist 2—Lena also reports from the same crime desk. She quite accepts the newspaper policy that stories should not negatively affect community relations and entirely trusts the newspaper owner's judgment as to whether any given story will do so. The penalty for transgressing the policy is immediate dismissal, which Lena regards as vastly disproportionate for what are often reasonable mistakes. She has a story that she does not believe will damage community relations but refrains from publishing because she expects the newspaper owner will judge otherwise.

Though Lena considers the consequences for publication unreasonable, it is not fear of those consequences that motivates Lena's silence. As such, her self-censorship is not pernicious as it is based on accepting the authority of the speech rule (and its application) to which she adheres.

On reflection, the different ways in which judgments of reasonableness bear on perniciousness under 2 and 3 respectively make a great deal of sense. Under 2, it's helpful to think about

parallels with coercion. Consider a classic "your money or your life" threat case. Structurally, this is like *Hostile Dinner Party*. Nate avoids speaking (hands over his money) in order to avoid negative self-affecting consequences—inviting a hostile reaction from others (losing his life). Now consider a variant of that classic case: "your money or I kill your best friend." The mere fact that the consequences one aims to avoid are other-affecting does not dissolve the coercive character of the threat. I hand over my money for first order moral reasons—to avoid risk to my friend's life—but I still accede to terms of conduct coercively impose by others. This is more like *Birthday Celebration*—Nate has first order moral reasons to refrain from speech that are triggered because of others' conduct that he regards as unreasonable fetters on the speech he would otherwise make.

Under 3, in contrast, the unreasonableness of the consequences doesn't have a direct bearing on perniciousness, because those consequences don't trigger the agent's reasons to respect the authority of the speech rule (in contrast to 2, where the consequences trigger the moral considerations). Unreasonableness has a bearing when the agent self-censors to avoid consequences that aren't reasonably entailed by the speech rule that they affirm—but this is really just a special case of the agent not recognizing the authority of *that* rule. They acquiesce to a standard of appropriate speech that isn't covered by the speech rule whose authority they grant.

In summary, the perceived reasonableness or unreasonableness of the costs or rule bears on the perniciousness, or otherwise, of self-censorship involving either first order moral reasons or affirmation of an authoritative speech rule but in different ways. In the former, self-censorship remains pernicious when the moral reasons are triggered by the consequences of others' judgments deemed unreasonable by the self-censoring party. In the latter, inasmuch as self-censorship aims to avoid an unreasonable application of a speech rule, it is pernicious because the agent does not in fact accept the authority of the speech rule on whose basis they fear being penalized.

One context in which self-censorship has been especially decried as a malignancy is higher education [6, 27]. It is instructive to briefly consider what this analysis reveals for self-censorship here. Consider:

Sanitized lecture—Otis drafts his lecture on the history of colonialism and concludes that he should refrain from including some of the details of certain atrocities in case they disturb his students too greatly.

Though the pedagogical merits of Otis's decision are open to debate, *Sanitized lecture* is a case of benign self-censorship, given that Otis endorses second order reasons to refrain from upsetting his students and does not acquiesce to their judgments. Compare:

Trepidatious lecture—Otis drafts his lecture on colonialism but is concerned that if he refers to literature outlining some of the positive effects of colonialism, his student evaluations will suffer because attendees will regard him as downplaying the injustices it entailed. He avoids referring to the literature in question.

Fearful lecture—Otis believes that the positive effects of colonialism are perennially underplayed and would ideally give a lecture that, in his view, corrects that picture. He fears, though, that his students will see him as a racist defender of past atrocities and engage in campus protests against his teaching.

Both *Trepidatious lecture* and *Fearful lecture* involve acquiescence. Otis refrains from speaking as he judges he ideally should, in order to avoid consequences commissioned by judgments of others. On this basis, both cases are, on my account, to some degree pernicious. However, their degree of perniciousness varies in line with the degree to which Otis acquiesces to judgments deemed not just incorrect but illegitimate as considerations to determine the bounds of acceptable speech or belief. To that degree he is further alienated from the standards of speech to which he reluctantly accedes.

5 | Self-Censorship, Moral Autonomy, and Authenticity

A distinctive vice of self-censorship, then, entails acquiescing to others' judgments concerning acceptable speech or belief. But why is acquiescence pernicious? It is pernicious because the person who self-censors abdicates their moral autonomy over speech. They restrain their speech not on the basis of considerations they independently judge significant, but to escape consequences of others' judgments on the appropriateness of their speech or belief. There are echoes here of R P Wolff's famous objection to state authority on grounds of its incompatibility with moral autonomy [28]. Still, there are also crucial differences. Wolff intimates that submitting to state authority involves ceasing to exercise one's own judgment, but self-censorship does not entail this [28], pp. 9–11. Rather, the form of abdicating moral autonomy I have in mind involves persons exercising judgment but taking the opposing judgments of others (or consequences of them) as decisive reasons against speech. Moreover, on Wolff's view a person abdicates moral autonomy when they accept the authority of others to enact moral requirement; however, self-censorship (of this form) precisely does not accept the authority of others. It involves conforming with but not embracing the judgments of others or their standing to determine acceptable speech or belief.

The contrast between principled silence and this self-censorship is crucial, then, when it comes to moral autonomy. When a loved one asks, "how do I look?" we often rightly guard more honest assessments in case an open book hurts more than it helps. Yet this noble restraint isn't a matter of *acquiescing* to their judgment or feelings—it is to take account of the impact of speech on others, mediated through their own view of the world. Matters are quite different when the eyes of others are ballasts that weigh on the decision to speak. Then the person's silence is born of fear of transgressing a form of external authority. This doesn't entail that others in fact have any authority, *de facto* or normative; formal or informal. Nor that others are judged as having normative authority—such as in *Schoolteacher*. Rather, that the self-censoring party acts as if others' judgments have that authority despite implicitly rejecting it. Here the self-censor, in effect, adheres to an alien standard of acceptable speech or belief on the basis of which penalties are exacted. Such expressive restraint

involves submission to boundaries of acceptable belief or disclosure dictated by others' judgments and not one's own. Their view of the world is dominant in shaping what is or is not said, not because the person accepts their authority as fitting, but to avoid ramifications of trespassing against it. Acting in ways that are subservient to social pressure of this kind is to conduct oneself as if one is not to be heard; not merely living an insular expressive existence but having that insularity formatively (albeit not irresistibly) foisted upon oneself by others.

So, a distinctive source of perniciousness in self-censorship is driven by acquiescence to others' judgments. This involves abdicating moral autonomy by adhering to standards of speech and belief set by others. This is not to say that acquiescence embodies a unique and full explanation of the badness of any and every troubling instance of self-censorship, not least because the perniciousness of acquiescence itself varies with other considerations, as I explain below. Rather, it constitutes a core and overlooked pernicious-making dimension. I have charted the course towards that conclusion through exploring a range of cases of self-censorship and contending that this account explains our sense of their respective benign or pernicious character. It is worth briefly seeing that acquiescence offers a more illuminating framework explaining the ills of self-censorship than alternative accounts. Those accounts either find it harder to accommodate the intuitive pernicious and/or benign character of certain cases, or else their root explanation of perniciousness ultimately seems undergirded by acquiescence.

First, as I have already indicated, this account better explains what's troubling about self-censorship than purely externalist accounts. For instance, a domination-based account might hold that self-censorship is pernicious when prompted by the threat of arbitrary interference and benign when not.[12] But this implies counterintuitive conclusions in a number of cases. As we saw in *Ideological Interview*, self-censorship can be troubling even if a person is not in fact dominated but simply *reasonably fears* arbitrary interference. A domination account struggles to explain this. Moreover, there are other cases that seem somewhat troubling even when the feared interference isn't obviously *arbitrary*. Both *Friendships* and *Abortion* entail self-censorship that is troubling to some degree, but neither involves fear of *arbitrary* interference, but reasonable reactions of others that together coalesce into social costs we might have reason to fear. Similarly, in *Reluctant Teacher* the troubling character of the self-censorship does not hinge on whether the penalties on speech are arbitrary or not. Conversely, a domination account suggests the wrong conclusion in certain benign cases. In *Confession*—involving self-censorship to avoid revealing incriminating facts—the self-censorship itself seems benign even if Sally is vulnerable to arbitrary punishment for her crime. At bottom, externalist accounts have difficulty explaining such cases because they locate the perniciousness of self-censorship in the circumstances that provoke it rather than attending to the reasons for which a person refrains from speech. Now, the distance between external factors and the ills of self-censorship should not be overstated—relations of domination and social suppression are precisely those liable to encourage pernicious forms, as I will explore below. In that way, externalist accounts can readily capture some of what is bad about situations in which persons are encouraged to self-censor. Even so, many cases of self-censorship are not pernicious primarily *because* it is the product of such relations but for reasons internal to the decision not to speak.

On my account, the perniciousness of self-censorship does not depend on the content or character of the curtailed speech but on the reasons for curtailment. This might be challenged. Perhaps pernicious self-censorship curtails valuable speech and benign curtails disvaluable. This would be a substantive, normative diagnosis of the ills of self-censorship based on the type of speech. I doubt this could support a compelling account, however. For one thing, it suggests that our judgments on the perniciousness of self-censorship depend on prior judgments on substantive, contested matters. Although the value of speech isn't necessarily exhausted by its content (its character and effects also matter), that content still has substantial bearing (the value of pro or anti-abortion speech partly depends on whether the propagated views are true or not). So we first need to establish the merits of gun control (*Hostile Dinner Party*), abortion (*Abortion*), religious beliefs (*Friendships*), sexual morality (*Reluctant Teacher*), etc. in order to sort benign from pernicious cases. Yet the intuitive badness of such cases does not seem to substantially depend on those judgments. In *Hostile Dinner Party*, for instance, Nate's self-censorship would remain just as pernicious, as self-censorship, if the tables were turned, and he was inclined to speak in favor of gun control and other guests were steadfastly against it. Moreover, such an account struggles to account for the intuitive difference between *Dinner Party* and *Hostile Dinner Party*. There is no difference in the value of the curtailed speech—it is the same—and yet the latter is intuitively pernicious in a way that the former is not. Of course, the all-things-considered badness or goodness of self-censorship will sometimes turn partly on the value of the speech (a matter to which I will return), but this does not entail that its character as pernicious or benign is necessarily explicable solely on such grounds.

Alternatively, one might accept that persons' reasons for self-censorship bear on its perniciousness, but still deny that abdication of moral autonomy explains this. One might, for instance, suggest that self-censorship is pernicious because it involves a different vice. Given that self-censorship involves dissonance between persons' impetus to speak and their silence, perhaps self-censorship is pernicious when a form of hypocrisy. A paradigm case of hypocrisy involves condemning others for things one is guilty of oneself. Clearly, self-censorship does not itself fit this paradigm. Perhaps, though, under a broader conception of hypocrisy it might. On a broad definition, hypocrisy involves outward behavior that does not cohere either with professed commitments or inner conviction. Thus conceived, self-censorship might be seen as a form of hypocrisy. One's inner convictions—about what to say—are at odds with one's outward behavior—what one in fact says. However, this does not supply an alternative explanation of the badness of self-censorship. Both benign and pernicious cases of self-censorship seem to involve hypocrisy, thus understood.

But perhaps the notion of hypocrisy can sort benign from pernicious cases in the following way—when a person endorses reasons to refrain from speech (first order moral ones, second order authority acceptance), then their inner convictions do in fact cohere with their outward behavior. This delivers the right

conclusion for some cases. It sorts *Dinner Party* (benign) from *Hostile Dinner Party* (pernicious), and *Schoolteacher (benign)* from *Reluctant Teacher (pernicious)*, for instance. However, there are reasons to doubt it is an adequate alternative to my abdication of moral autonomy account.

First, it's not clear that this conception of hypocrisy delivers a root explanation of perniciousness. If avoiding hypocrisy requires coherence, we still need to know what's important about coherence. Second, this conception of hypocrisy trades on a specific understanding of coherence. Convictions cohere with behavior when the agent endorses certain reasons for that behavior. But this is less an alternative to my account and more a repackaging of one of its key components. Third (because that repackaging is only partial), this account struggles to explain why some cases are intuitively benign despite a lack of coherence. In both *Bargaining* and *Confession*, for instance, the self-censorship lacks this form of coherence (the parties keep silent for prudential reasons, rather than for moral or authoritative ones), but neither seems troubling as cases of self-censorship.

Given that the badness of hypocrisy itself demands explanation, perhaps some such explanation can offer a superior account of the ills of self-censorship. Three natural places to look are to complicity, integrity, and authenticity. One diagnosis is that certain cases of self-censorship are pernicious when they entail a certain type of complicity. Perhaps failing to speak in opposition to others' views involves complicit support for views one judges false or objectionable. This might make sense of Nate's self-censorship in *Hostile Dinner Party* insofar as he does not contradict the dominant pro-gun control tenor of the conversation.

However, as a general explanation of pernicious self-censorship, this faces several disadvantages. First, silence only amounts to complicity (if at all) in specific contexts, and those contexts don't necessarily mirror pernicious cases. The most compelling case for silence-as-complicity applies to instances where one has (or takes oneself to have) a duty to speak out against conduct or beliefs of others. However, there are many pernicious cases of self-censorship where the agent doesn't regard themselves under a duty to express something. *Ideological interview* is just one such case, where Mariane curtails her speech to avoid being perceived as a socialist but does not take herself to have a duty to express socialist commitments, still less in the specific context of the interview.

A slightly different case for silence-as-complicity, though, is that remaining silent can itself express support for others' views. In certain contexts, remaining silent plausibly expresses agreement with what others suggest or assert [30, 31]. Indeed, this is the most compelling explanation of any would-be complicity in *Hostile Dinner Party*—Nate's lack of dissent in the pro-gun control conversation expresses agreement, or so the thought goes. Perhaps, then, the perniciousness of self-censorship is explained by complicity of this kind—expressive support for speech or belief with which one disagrees. This shares with my account an emphasis on dissonance between the judgments of self-censoring parties and those of others. The drawback of this approach, though, is that perniciousness is often not plausibly contingent on the expressive force of silence. For one thing, silence does not always express agreement in the face of others' assertions [32]. Nate's silence could be reasonably interpreted as indifference, or politeness, rather than agreement with the pro-gun control tone of the dinner party discussion. It would remain to some degree pernicious even then. Moreover, pernicious self-censorship need not involve silence in the face of opposing assertions. Suppose that the conversation is not about gun control specifically but involves an invitation to raise political issues that inform one's voting decisions. If Nate has good reason to think that expressing an anti-gun control position will invite social penalty, his self-censorship looks pernicious but hardly expresses agreement with the opposing view. Or suppose Clarissa, in *Abortion*, is considering whether to attend a pro-choice demonstration but fears the consequences of doing so. Her staying at home doesn't express support for the opposing view but remains troubling nonetheless.

Instead, we might turn to a cousin of complicity and explore whether *integrity* will better explain perniciousness in self-censorship. This is closer to the mark, and correctly spotlights the misalignment between the self-censoring party's judgments and their reasons for silence. Moreover, I quite agree that integrity bears on the degree of perniciousness in certain cases of self-censorship. I doubt, though, that it is the base explanation for their perniciousness. On a broad understanding of integrity, it involves acting in ways incongruent with one's values or beliefs. So understood, for self-censorship to compromise integrity it is not sufficient that the silence conceals one's values or beliefs; it must in some sense contradict them. Simply refraining from expressing one's beliefs is not necessarily integrity-compromising; then, one must also be committed to valuing the expression of those beliefs. Some beliefs are like this. A person may, for instance, have religious convictions that themselves entail testifying to the truth of those convictions. If that person remains silent, then their actions are incongruent with their commitments in an integrity-compromising way. But not all beliefs take this form. In *Hostile Dinner Party*, Nate may well be inclined to voice his views about gun control (though he ultimately acquiesces), but this need not involve anything akin to a commitment to generally doing so. Despite being pernicious, it isn't clear that Nate's self-censorship is integrity compromising. Of course, it's true in this case that Nate refrains from saying what he otherwise might. In that sense, he doesn't live in accordance with his projects. But merely being frustrated in living in accordance with one's values does not itself amount to integrity damage.

What is more, some benign cases plausibly involve sacrificing integrity. In *Schoolteacher*, suppose that Jeannette has an ongoing and deep commitment to espousing creationism but accepts her role teaching evolution in order to earn more money. Even if she affirms the authority of the school to determine classroom speech, she plausibly compromises her integrity.[13] There is perniciousness here—because of the integrity compromise—but it isn't the self-censorship itself that is pernicious.

Finally, it might be argued that the root of perniciousness in self-censorship is inauthenticity. Again, this is surely a part of the explanation for more pernicious cases. However, such an alternative struggles to differentiate between principled and fearful cases of self-censorship. In both *Dinner Party* and *Hostile Dinner Party* Nate conceals his authentic self, but the reasons that motivate his doing so clearly bear on whether this is pernicious.

Moreover, some cases look pernicious even when they involve curtailing *inauthentic* speech. When a stand-up comedian carefully edits their routine to avoid social backlash, they don't necessarily curtail their authentic self, but the self-censorship seems to some degree pernicious even so. It could be argued that I am construing inauthenticity imprecisely—to encompass cases where one hides some truth about oneself. One might suggest, in contrast, that self-concealment or misrepresentation can itself be authentic, where this reflects the person one is. The comedian may not believe the things they say on stage—but the persona they adopt is still, in a certain sense, true to their self. Perhaps, then, authenticity involves acting for reasons that cohere with one's own sense of what is important or valuable, and sometimes this will involve judgments about how much of one's beliefs to expose to others. This might distinguish *Dinner Party* from *Hostile Dinner Party* where, in the former, Nate acts authentically because he judges silence appropriate. This seems close to the truth; however, now this seems to simply redescribe an exercise of moral autonomy. Namely, that one affirms silence for reasons one endorses, even if it involves resisting the inclination to disclose one's beliefs and values to others.

Although neither authenticity nor integrity is the root of perniciousness in self-censorship, they remain important factors that accentuate or diminish perniciousness, whose origin lies elsewhere. That perniciousness is underpinned, ultimately, by the abdication of moral autonomy. This embodies a form of expressive dependence—what I am willing to say is dependent on the judgments of others. Still, this is most deeply troubling when it pertains to *authentic* self-expression.

The closest ally of my suggestion here is the thinker-based defense of free speech from Shiffrin. At the core of her view is our interests in being known by others and in arbitrating, for our own reasons, what to reveal of ourselves to others. Shiffrin writes, "what makes one a distinctive individual *qua person* is *largely* a matter of the contents of one's mind, to be known by others requires the ability to transmit the contents of one's mind to others ... Communication of the contents of one's mind ... uniquely furthers the interest in being known by others." [33], p. 89.

Authenticity thereby embodies a further consideration that bears on the intuitive perniciousness of different instances of self-censorship. Consider a bookstore owner declining to stock work by an infamous author, not because the owner themselves deems the work objectionable but because of public pressure against doing so. We should be unsettled by this, but not as disturbed as when fear of social judgment stops the author from writing at all. Or take a class of students debating a politically sensitive question. We should regret if none are willing to play devil's advocate and float a controversial stance for fear of classmates' scorn and contempt. Apart from the deferential dynamic this embodies, the discussion might lack vigor and abrasiveness so often useful for fruitful learning. Yet it is still unhappier if the withheld views are actually held by the students who keep quiet. Those students conceal who they are and what they believe in because of the threat of peer disapproval. The perniciousness of authentic acquiescence is further amplified by the centrality of those beliefs for the individuals who acquiesce to conceal them. Here, then, integrity has a place in accentuating the perniciousness of self-censorship. Withholding one's views on history's best basketball player is substantially different from hiding one's religious beliefs—it matters how integral the conviction is to one's self-understanding.

At its most troubling, then, acquiescent self-censorship involves surrendering control over how far we are known to others. I refrain from disclosing elements of myself, not for reasons that I judge independently authoritative, but in order to evade the judgments (and attendant consequences) that would form if others were to see who I really am. It involves sacrificing the basic interest in being known to others on the altar of others' judgments of who one should be or what one should say.

6 | Political Correctness, Deliberative Freedom, and the Soil for Self-Censorship

Abdication of moral autonomy over speech is the root of what is bad about much self-censorship. Not necessarily its lone defect, but a central and regrettable one. On this account, the perniciousness of self-censorship derives from the reasons why persons restrain their speech. In that sense, I have thus far offered a primarily internalist diagnosis of self-censorship—as an individual vice that can be comprehended without essential reference to the external conditions that give rise to it. Self-censorship is, however, rightly decried as a symptom of a deeper malaise—the propensity of society to sanction and suppress speech. "Cancel culture" is the fashionable phrase of today, but in the past this, or something similar, flew under the banner of "political correctness" [34, 35]. Even among those who lament the onward march of repressive attitudes to free speech in recent years, there are contrasting attitudes to political correctness [34, 36]. Some regard the time of political correctness as a halcyon era, engendering respectful discourse and sanding the edges of more abrasive language before regrettably morphing into distorted pathology [37]. Others see seeds of tyranny sown from the start [38]. A milder cousin of the modern expressive despotism: less muscular and erratic than cancel culture run amok but with the same proclivities nonetheless.

There is a sensible and persistent worry, then, about a culture hostile to expressing certain things that, in turn, breeds a tendency towards self-censorship. Though I earlier challenged purely externalist accounts of the ills of self-censorship, the social conditions that produce fertile soil in which self-censorship can fester remain an integral part of the story. In relation to this, there are three important lessons to be drawn from the account of self-censorship as acquiescence. First, it allows us to better diagnose any reasons we have to lament political correctness. Second, in relation to general free speech concerns connected with self-censorship, the account indicates that we have reason to care about individual cases of self-censorship beyond any negative systemic chilling effects on speech. Third, and most importantly, the account illuminates particular ways in which social penalties on speech are distinctively concerning.

First, then, on political correctness, the account helps us appraise the character of political correctness, at least in relation to self-censorship. It reveals that our reason to lament political correctness depends substantially on the impetus for a person's

self-censorship. For the true believer, political correctness is a form of principled silence under which the tenets of putatively respectful speech are positively endorsed. Less doctrinally immersed is the willing disciple, who doesn't comprehend exactly why a turn of phrase is morally suspect but accepts the current social more against it being said. The case we should be worried about here, though, is that of the fearful subservient, who regards social speech codes as nonsense but dares not overstep them. A key reason we have to regret an oft-maligned culture of political correctness, then, is not that it entails prescriptions for appropriate speech as such—even very strict ones—but when people comply with those strictures for fear of transgressing the standards of others, rather than endorsing those speech rules as their own.

Second, the account also tells us something about the character of self-censorship as a free speech concern. As I mentioned at the outset, while self-censorship is ubiquitously posed as a free speech-type issue, its bearing here is not transparent. Particularly given that self-censorship entails voluntary expressive restraint, it is not obvious how *freedom* of speech is imperiled. One line of thought involves pointing to common values served by freedom of speech that are also endangered by a climate of self-censorship. The common currency between self-censorship and censorship proper is that something goes unsaid. The broader values served by such speech seem to give us reason to care about self-censorship in the same way we have reason to care about censorship in general.

A Millian view that restrictions on speech dull citizens' intellectual engagement and hinder progress toward truth has an evident purchase here [39], ch. 2. Just as when people are subject to external censorship, if people self-censor, they don't deliver ideas into the world that they otherwise would. Perhaps this results in fewer conflicts between contrasting opinions, where truth is most expeditiously approached, as Mill tells us, "by the rough process of a struggle between combatants fighting under hostile banners." [39], p. 45. A close relative of the Millian case for freedom of speech is a specifically democratic worry—where limits on free speech frustrate hearer autonomy because they shear away different perspectives and undercut independent formation of belief. This also nurtures epistemic indolence wherein social conformity and received opinion tend to thrive. Democratic opinion is formed not by separate, autonomous individuals exchanging, assessing and critiquing in an open, unregulated way, but by being tilted towards some conclusions and away from others. As Meiklejohn puts it in his influential defense of free speech on this basis, the danger of state regulation of speech is that it can involve "the mutilation of the thinking process of the community." [40], p. 26. Though this is often framed in terms of an objection to state control of opinion specifically, it's clear that the worry might also pertain to self-censorship to some extent as well.

Both the Millian and democratic cases, thus understood, regard what we are deprived of as *hearers*—denied a variation of perspectives as those who might impart them fear treading outside of the range of acceptable (and, crucially, unpunished) beliefs. These hearer-based concerns are echoed in contemporary work on the epistemic implications of excluding certain perspectives from the discursive domain. Joshi, for instance, argues that social pressure against expressing a certain belief can produce a lop-sided evidence set concerning the merits of that belief [41]. This dynamic renders it more likely that epistemic bubbles will emerge, within which relevant perspectives have been left out [42]. My account dovetails with these hearer-based concerns about epistemic health [43], but also adds a distinctive speaker-based reason. It complements such accounts inasmuch it is precisely dissident and heretical views that people are liable to hide from others. So acquiescent self-censorship tends to temper the discursive friction conducive to epistemic value [44, 45]. However, the same is true regardless of our reasons for not speaking. Principled silence equally involves withholding thoughts from other people. Or fellow citizens might just be of like mind with one another. These hearer-based accounts that tie self-censorship with wider free speech concerns thereby struggle to identify anything distinctive about self-censorship that makes it any more troubling than principled silence, excessive reticence or passivity.

What is more, though independent belief-formation seems threatened when full blown censorship restricts access to others' points of view, it is less clear that it is undercut when others voluntarily refrain from imparting their view into the discursive domain. Sure, we lack some knowledge about what others think, but this does not necessarily, of itself, threaten our autonomy as hearers.

Added to this, the Millian and democratic accounts point towards systemic failures in the discursive scheme, rather than anything troubling about individual cases where some refrain from speech. Likewise, work on the so-called chilling effect of laws and norms often has a similar systemic bent—whereby a central worry is the general effect of broad or vague laws on social discourse [46].[14] Provided social discourse remains in relative health, these accounts fail to explain any would-be perniciousness concerning individual cases of self-censorship. Indeed, Meiklejohn is explicit on this when he argues that "what is essential is not that everyone shall speak, but that everything worth saying shall be said." [40], p. 24. Self-censorship as acquiescence offers just such an explanation. Beyond any chilling effect on wider social discourse, there is something inherently pernicious about acquiescing to alien standards of acceptable belief or speech, in such a way that a person relinquishes moral autonomy and, in turn, forgoes a more independent determination of what to express.

This also helps us more fully understand what is freedom-threatening about social milieux hostile to certain kinds of speech, even when that does not tilt into full blown societal censorship. When non-coercive but systemic antagonism towards heterodox belief fosters self-censorship, it discourages expressive independence. Though social pressure does not strictly censor persons' speech, it does potentially muffle the autonomy of speech by presenting a set of extraneous considerations at odds with persons' own sense of acceptable speech or belief. Not only is social sanction liable to encourage an abdication of moral autonomy, then, it also poses a broader threat to deliberative freedom. Such freedom is hampered when deliberation over what to say and believe is colored by considerations that ought, ideally, to be extraneous to the matter. Shiffrin offers a perceptive account of this danger. Persons have an interest, she argues, in

"revealing, sharing, and considering ... mental contents largely at their discretion, at the time at which those contents seem to them correct, apt, or representative of themselves, as well as to whom (and at that time) such revelations and the relationship they forge seem appropriate or desirable." [33], pp. 78–88.

In self-censorship as acquiescence, it is precisely for reasons whose normative authority I deny that I refrain from speech. Where social sanctions encourage that self-censorship, they burden expressive deliberation by cluttering the process of weighing and deciding whether to speak and why—tempting us to accede to their force even while we deny that we should. Even when I don't actually self-censor, then, my deliberation as to what to say is tugged and pulled by considerations that, on some level, I deem inapt. Illustrations of this risk to deliberative freedom abound. Spouses have reason not to react too angrily, even for legitimate grievances, lest the hostility figures too prominently in the deliberations of the marital partner, distracting from its cause and supplanting proper motivation to alter future behavior for the right reasons. Likewise, parents have reason to refrain from chiding children too harshly, in case fear of punishment overawes the child's reflection on the transgressive behavior itself. The core of the claim here, then, is that there is health when deliberation is less polluted by considerations to which it would ideally not need to be attentive [47], ch. 3.

In other work, Shiffrin illustrates the concern with distorting factors in relation to employment accommodation practices [48]. A paradigm case is Muslims being afforded breaks from usual working hours to hold Friday prayers. Without such exemptions the Muslim is forced, when deliberating whether and when to pray, to take into account the impact on her employment prospects. She must consider fidelity to religious commitment (faith, duty, identity, service, etc.) in the light of, and in competition with, prudential reasons of occupational choice and income stability. As social sanctions stalk a person's evaluation over what to say or think, her deliberative freedom is increasingly cloistered from reasons she judges central, and she is further nudged away from beliefs she would express in conditions of greater independence. Even where social penalties on speech and belief do not amount to restrictions on the freedom to speak, then, they hinder deliberative freedom by fettering the negotiation between reasons concerning what to say and potentially steering persons towards acquiescence.

The third, and most important, counsel to be drawn from self-censorship as acquiescence is an adjudication over the types of social penalties on speech that we should be concerned about. I began by noting that we routinely invite social costs from others. This self-evident feature of social life forms the bedrock of many a retort from those who downplay the sometimes-capricious ferocity of societal punishments on speech. The retort is only partly astray. Sometimes social consequences of speech are rightly ours to bear. At others, however, as Mill forewarned, they fume from the "engines of moral repression" [39], p. 17. The justice of social consequences on speech is vast question, with horizons far beyond my ambition here. In broad strokes, though, my claims about self-censorship suggest something of note. Namely, that sanctions potentially justifiable to their victims are less likely to manifest the ills I have lamented. Just as acquiescent self-censorship yields to consequences rooted in others' judgments, so deliberative freedom is imperiled by an expressive environment hostile towards, and punitive of, heterodox beliefs—exacting penalties based on a standard of acceptable speech alien to some of its subjects. Far less of a threat is posed by social sanctions that can be readily justified to the person to whom they apply. When a person can either see the legitimacy of the standard against which they are being judged, or when they share the judgments themselves, both acquiescence and deliberative unfreedom are less likely results.

Ceteris paribus, then, sanctions (reasonably perceived to be) based in judgments that a person's belief is false or objectionable are more likely to encourage acquiescent self-censorship and burden deliberative freedom. Exacting penalties on a person's speech because their expressed beliefs are objectionable means sanctioning based on first order reasons to which the sanctioned person cannot in good faith assent. There are, however, two ways that the sanctioned person might affirm the legitimacy of the sanctions nonetheless.

First, the person can more readily view them legitimate if the *consequences* of speech are the targets of sanction, rather than the expressed beliefs themselves. Suppose Sally leaflets her local area with anti-abortion flyers. Doing so risks vilification and ostracism from others in the local community. Such sanctions can be motivated by very different concerns. The neighbors might deem Sally's speech objectionable because they believe it embodies an immoral lack of regard for women's bodily autonomy. Alternatively, perhaps a spate of local abortion clinic bombings means the neighbors judge Sally's speech irresponsible, quite apart from their thoughts about her views themselves. The latter is accessible for Sally where the former is not. In principle, she can affirm reason to refrain from leafleting, still resolute in her beliefs about abortion. But for the former, Sally declining to express her views involves shrinking from their inimical gaze. The neighbors' punishment has a basis that necessarily escapes being shared by Sally herself.

Second, sanctions marshaled in defense of weighty associative interests are more seamlessly apprehended by the socially punished as legitimately imposed. Say Malachi disagrees with the editorial direction of the newspaper for which he works, but sees the worth of collective message discipline, and understands being penalized for publicly straying from the party line. Or recall Ideological Interview. Mariane self-censors to avoid consequence rooted in antipathy towards social sympathy. But suppose she is interviewing for a vehemently capitalist political party. She might still conceal her convictions but think the grounds for exclusion fair enough.

Together, these observations suggest that specifically punitive social sanctions targeting people for beliefs offer particularly tender ground for self-censorship and the erosion of deliberative freedom. Less where the borders of associations are preserved by excluding dissident voices, but where there is a Salem-esque appetite for seeking heretics to disavow and punish. It is as much the *location* of social sanctions as their motivation that matters here. Above I posed the example of Muslim prayer as an illustration of deliberative freedom under threat. It is a central case precisely because it concerns fundamental commitments fettered by common inconveniences—the mundane encroaching on the sacred by taking

undue prominence. But the converse paradigm also applies—where matters of conscience and conviction intrude into more ordinary areas of life, lacing fraught and contested matters through everyday activities in which they don't have a natural bearing, shadowing our decisions therein. Where you are unwelcome to sit in the stands of the big game unless you worship the right god, obstructed when establishing a business unless you hold the right political views, prevented from taking out a bank loan unless you hold the right moral beliefs. The virtue of toleration is as much about untethering these irregular bedfellows as anything else.

These general reflections concerning the forms of social sanction liable to foster self-censorship also commend institutional antidotes. My account supports regulatory measures in order to limit the scope for societal animus against unpopular belief translating into formal associative sanctions that would render the social costs on speech and belief more significant and pervasive. Rights against belief-based discrimination in certain associations, including workplaces, are a central case. Limiting employer rights to discriminate against employees on belief-based grounds insulates individuals from social costs grounded in beliefs they reject. In turn, awareness of those protections helps assure individuals that they can speak without fear of this specific source of sanction. Of course, associations often have strong interests in discriminating on belief-based grounds. Paradigmatically, religious organizations have interests in hiring for leadership positions based on religious belief.[15] Such cases are, however, precisely those where sanctions are more likely to be judged legitimate, given that they serve strong associative interests. In broad strokes, then, this commends i. clear bases on which employees can claim against belief-based discrimination, ii. robust, low burden procedures for pursuing claims, iii. transparent hiring criteria compatible with protections against belief-based discrimination, and iv. perspicuous carve-outs licensing belief-based discrimination grounded in strong associative interests. As I allude to above, pernicious self-censorship can be driven by beliefs and convictions taking undue salience across certain spheres of everyday life. This also points towards the need for a high burden on associations to demonstrate that extra-mural speech threatens those associative interests.

In an academic context, beyond the belief-based protections noted above, my account commends insulating teachers from sanctions for the content of teaching, notwithstanding the need for clear requirements to teach material with disciplinary relevance, and the duty of care to avoid harassment, discrimination or stigmatization. More broadly, it implies the desirability of clear guidelines for appropriate classroom speech that are i. consistent with the central place of academic freedom and ii. discourage informal sanctions on speech within those bounds. An explicit framework demarcating the limits of appropriate classroom speech, justified on grounds of academic freedom and pedagogical value, supplies participants with a standard of acceptable speech whose authority they can accept, while also advertising the appropriateness of other expressive contributions.

7 | Objections

Now, although my account sheds light on certain bad-making features of a speech-punitive culture, the perniciousness of self-censorship as acquiescence still has an essentially internalist character. It is a matter of the reasons for which people refrain from speech. It might be argued that an internalist view is incapable of explaining the distinctive *wrongfulness* implicated in self-censorship. After all, it's not clear whom we are wronging by simply manifesting the vice of abdicating expressive independence. My account thereby faces a hurdle more seamlessly overcome by externalist counterparts. One route around this is to advocate for a positive responsibility to speak that self-censorship would transgress [44].[16] Ronald Dworkin, for instance, forges this line by defending a vision of ethical individualism under which persons have "a duty to speak out for what one believes to be true." [51], p. 188. There is something compelling here, however any such duty to speak the truth will evidently be heavily qualified, given that there are innumerable occasions in which not speaking the truth is morally permitted or required. As we have seen, many cases of self-censorship are benign. So any such explanation for the wrong in self-censorship will depend on identifying what about yielding to others' judgments through acquiescence is specifically wrongful.

A more fruitful avenue for explaining such wrongfulness, however, highlights our role as participants in the replication of social conditions that foster acquiescent self-censorship. When we uphold a culture that tends towards the infliction of penalties encouraging such self-censorship, we risk complicity in a system that discourages expressive independence. Sufficiently formidable social sanctions suppress speech in a thoroughgoing fashion, robbing people of expressive independence and wronging them overtly. However, even less overbearing social consequences—that leave the freedom to speak sufficiently intact—can still engender a climate where persons are encouraged to cower from negative social consequences born of transgressing norms of acceptable speech and belief held by others as the operative reasons to refrain from speech. We wrong others when we foreseeably and avoidably contribute to a discursive culture with such effect.[17] Now, these are admittedly cursory remarks, with laden terms, that engage the much broader issue of individual responsibility for manifesting collective harm. The point is, though, that the internalist character of self-censorship does not necessarily sever it from questions of wrongdoing. Rather, understanding the perniciousness of certain forms of self-censorship helps us more fully comprehend the type of discursive culture we have a responsibility to refrain from being complicit with.

The most natural objection to my account is that even acquiescent self-censorship is sometimes good. Witness the sincere racist who refrains from racist diatribes for fear of social repudiation and ostracism. Surely, a world with such self-censorship is preferable to one where unrestrained hostility and vitriol are let loose. Far from being a morally troubling phenomenon, then, self-censorship is a crucial element of the good society. While there is a sage lesson here, it isn't one which casts doubt on my account. It is true that self-censorship can be all-things-considered good. But even in cases where it is good that someone self-censors, it remains *pro tanto* regrettable when they are motivated by a fear of social consequences issued by judgments they don't share, rather than convictions of their own. Better for people to refrain from racist diatribes because they can see the wrong, rather than from fear of social punishment.

This rejoinder will not satisfy some. Surely, the worry goes, there are some abhorrent beliefs about which there is nothing regrettable when they go unexpressed, even where acquiescent self-censorship is implicated. Whatever the force of this concern, it can be accommodated on my account. This is because it is consistent with the general shape of my view to hold self-censorship is *conditionally* pernicious (say, only when it involves non-abhorrent beliefs or speech). Of course, this does not answer the broader question of how to balance the ills of self-censorship with the merits of withholding objectionable or harmful speech, but an answer to this relies on a broader account of the value and harms of speech. My ambition here is more circumspect—to articulate and explain a bad-making feature of self-censorship that has weight in that calculation. Now, for all that I have said decrying the perniciousness of acquiescence, some may argue that acquiescing to others' views of acceptable speech and belief is a natural and palatable practice necessary to grease the wheels of social cooperation. The truth in this is that some self-censorship has real social value, but it's not clear that acquiescent forms embody this. After all, if persons cower from ramifications for speech and belief they deem unreasonable, it's simply less clear that social cooperation is going well.

The above reservation concerning racist speech might be pressed in stronger terms, however. I proposed that the wrong connected with self-censorship lies with the cultivation of conditions that encourage it. The acquiescent sincere racist embodies an instance in which it is good that something isn't said, even if there is an element of perniciousness. If so, it's hard to see how upholding conditions that foster acquiescent-but-good self-censorship involves complicit wrongdoing. After all, a society without racist speech is precisely one we should all aim to bring about. This is an important insight, but it doesn't follow that there is nothing wrong involved in upholding a speech-punitive culture that fosters acquiescent self-censorship. Rather, it helps us see that any account of the duties against complicity in encouraging self-censorship also relies on understanding the kinds of speech it is all-things-good to muzzle, weighed against the risk of producing a general culture of acquiescence for speech whose erosion isn't worth the price paid by a sacrifice of expressive independence. The very same question—of when the benefits of social sanction outweigh the costs—also applies when the issue is *societal* censorship, and the force of social penalty *restricts* rather than merely discourages speech. Where our duties lie in relation to these various aggregative speech-suppressing effects is an immensely complex broader question, but one that our present social and political culture suggests is ever-more salient and perhaps unlikely to fade in the near future. There is more work to be done here, but a clearer understanding of the distinctive perniciousness of self-censorship is an integral part of the picture concerning our responsibilities to sustain a suitably free and independent discursive culture.

8 | Conclusion

Self-censorship is a familiar notion and often central to worries concerning the suppression of speech yet used variously and often imprecisely. It should come as no surprise that self-censorship is elusive and contested. We can all agree that sometimes it is good to speak up and sometimes good to keep quiet. Naturally, though, people disagree over which cases are which. We disagree, moreover, about which pressures against speaking are benign and which are nefarious. My account identifies an important vice of a certain type of self-censorship that, in turn, helps more fully understand the ills of a society that inflicts punitive sanctions on persons holding and disclosing beliefs. The core of that vice is acquiescence to others' judgments concerning acceptable thought and expression. This involves relinquishing expressive independence and ceding the determination of bounds of appropriate speech to others. The notion of acquiescence illuminates why agents' reasons for restraining speech are important for understanding why some self-censorship is more troubling than others. It explains why deferring to others' views or rules on acceptable speech or belief isn't necessarily pernicious, for instance if we take ourselves to have first order moral reasons for doing so, or second order reasons to recognize the authority of the rules or judgments of others. Those cases are less concerning because they embody the exercise of one's moral autonomy over the determination of what to say, rather than its abdication. Recognizing the centrality of moral autonomy in turn allows us to discriminate more sagaciously between social penalties whose antagonism against otherwise voluntary speech might remain unclear. Those sanctions on speech and belief readily affirmed by the silent party as having a legitimate place—either as speech rules or reasonable reactions—are less liable to foster expressive submission and cumber deliberative freedom.

This is not to argue that acquiescence constitutes a necessary and full basis for perniciousness in any and all such forms of self-censorship. There may very well be further bad-making features of self-censorship cases of various kinds.[18] Still, my account identifies a basis for perniciousness in self-censorship that is at once more general and more foundational than alternatives. I have also contended, though, that the perniciousness of self-censorship is multidimensional and, moreover, a matter of degree, rather than being monolithically vicious or benign. Even acquiescent forms of self-censorship are less thoroughly troubling to the extent that silent parties recognize the legitimacy of the pressures on their speech. Though the self-censoring party doesn't share the views of acceptable speech or belief on which the feared social consequences are founded, they nonetheless apprehend them as reasonable responses of others. On the other hand, the reasons to regret self-censorship are still more acute when it is our authentic and central beliefs that we conceal for fear of others' opposing judgment. The more thoroughgoing that concealment is, the more completely it embodies estrangement from the discursive community fostered by others' alien judgments about the acceptability of our speech or belief.

Acknowledgments

Versions of the paper were presented at an MPSA Conference and a PPE Society meeting, and I'm grateful to audiences for their helpful feedback. Thank you to Jay Howard for detailed written comments and helpful discussion on the paper. I'm also very grateful to several anonymous reviewers, including two from this journal, as well as the editors of Philosophy and Public Affairs. Their constructive comments and advice prompted substantial improvements to the paper.

Funding

The author has nothing to report.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Endnotes

[1] Notable exceptions include [9–13].

[2] Some contrast self-censorship with state-censorship [14–16], or formal censorship [3] but others do not [17]. In that latter camp, some work associates self-censorship with the chilling effect of laws [18–20]. In some work, the notion of self-censorship enfolds a broad church of reasons for silence, including respect for others [21], or self-confidence [22], whereas other work ties self-censorship specifically with sanction-avoidance [23].

[3] For an account of reasonable alternatives in relation to "voluntariness," see [24], p. 139.

[4] Strictly, that they refrain from speech in order to avoid those unreasonable consequences.

[5] As I explain below, reasonable but mistaken judgments concerning the consequences of speech will similarly apply to self-censorship cases.

[6] For another view that emphasizes the relation between censorship and authority see [25].

[7] In more recent work Festenstein sketches three models of power—liberal, market-ideological, and Foucauldian—although this remains a broad strokes survey of different dynamics [11].

[8] Albeit for second order reasons to accept that authority, rather than for first order reasons affirming the content of the standards themselves.

[9] This may be poor judgment (or not!), or embody an unhealthy lack of intellectual confidence, but if these are vices, they are different from the focus of my account.

[10] Again, the distinctive perniciousness in *Biased reference* is not due to the power relations present here and absent in *Insecure graduate*. Both Aria and Julius are equally vulnerable.

[11] Or epistemic authority of the person whose beliefs they adhere to when self-censoring.

[12] The *locus classicus* of this view is [29].

[13] Or suppose Nate arrives at Dinner Party deeply committed to persuading others against gun control but refrains from doing so because he is can't bring himself to upset Matilda. It's not clear that this would make his self-censorship pernicious, but it tilts towards a sacrifice of integrity.

[14] As Simpson suggests, the metaphor of chilling itself connotes group-level concern as it "conjures and impression of chats becoming frosty. Fewer people want to speak, but also, where people do speak, things are less free-flowing. The dialogue freezes." [46], p. 8.

[15] For a relatively recent elaboration of such interests see [49], especially chapter 5.

[16] For a qualified critique see [50].

[17] Beyond a negative duty to refrain from inflicting certain social costs on speech, this might itself encompass a different positive duty to speak out; one grounded on a duty to counter a climate in which others fear to speak, rather than a duty to bear witness to perceived truths. Hannon alludes to a duty along such lines when he suggests we have a "duty to *make it less costly for others to share evidence*." [50], p. 286 (original emphasis).

[18] Indeed, we have alighted on some possible such vices in the course of the analysis. Perhaps Julius's deference to his mentor's judgment in *Insecure Graduate* betrays a regrettable lack of intellectual confidence, or perhaps Otis, in *Sanitized lecture*, shrinks from a responsibility not to coddle his students.

References

1. R. Adekoya, E. Kaufmann, and T. Simpson, "Academic Freedom in the UK: Protecting Viewpoint Diversity," Policy Exchange (2020), https://policyexchange.org.uk/wp-content/uploads/2022/10/Academic-freedom-in-the-UK.pdf.

2. E. Ekins, "Poll: 62% of Americans Say They Have Political Views They're Afraid to Share" (2020), https://www.cato.org/survey-reports/poll-62-americans-say-they-have-political-views-theyre-afraid-share#introduction.

3. J. Gibson and J. Sutherland, "Keeping Your Mouth Shut: Spiraling Self-Censorship in the United States," *Political Science Quarterly* 138 (2023): 361–376.

4. M. Goodwin, *Is Academic Freedom Under Threat?* (Legatum Institute, 2022).

5. A. Hegdal, *Freedom to Publish. Challenges, Violations and Countries of Concern* (International Publishers Association, 2020) 2020.

6. N. Honeycutt, S. T. Stevens, and E. Kaufmann, "The Academic Mind in 2022: What Faculty Think About Free Expression and Academic Freedom on Campus. The Foundation for Individual Rights and Expression" (2023), https://www.thefire.org/researchlearn/academic-mind-2022-what-faculty-think-about-free-expression-and-academic-freed.

7. E. Kaufmann, "'Academic Freedom in Crisis: Punishment, Political Discrimination, and Self-Censorship', Center for the Study of Partisanship and Ideology 2" (2021).

8. K. A. Naughton, "'Speaking Freely: What Students Think About Expression at American Colleges' FIRE Student Survey" (2017), https://www.thefire.org/research/publications/student-surveys/student-attitudes-free-speech-survey/org.

9. P. Cook and C. Heilmann, "Two Types of Self-Censorship: Public and Private," *Political Studies* 61 (2013): 178–196.

10. M. Festenstein, "Self-Censorship for Democrats," *European Journal of Political Theory* 17 (2018): 324–342.

11. M. Festenstein, "The Ethics and Politics of Self-Censorship," in *The Routledge Handbook of Philosophy and Media Ethics*, ed. C. Fox and J. Saunders (Routledge, 2023), 45–55.

12. J. Horton, "Self-Censorship," *Res Publica* 17 (2011): 91–106.

13. J. P. Messina, *Private Censorship* (Oxford University Press, 2023).

14. D. Bar-Tal, "Self-Censorship as a Socio-Political-Psychological Phenomenon: Conception and Research," *Advances in Political Psychology* 38 (2017): 37–65.

15. A. F. Hayes and J. Matthes, "'Self-Censorship, the Spiral of Silence, and Contemporary Political Communication," in *The Oxford Handbook of Political Communication* (Oxford University Press, 2017), 763–776.

16. K. Sharvit, D. Bar-Talb, B. Hameiri, et al., "Self-Censorship Orientation: Scale Development, Correlates and Outcomes," *Journal of Social and Political Psychology* 6 (2018): 331–363.

17. R. A. Sedler, "Self-Censorship and the First Amendment," *Notre Dame Journal of Ethics and Public Policy* 25 (2012): 13–45.

18. J. L. Bruneau, "Injury-In-Fact in Chilling Effect Challenges to Public University Speech Codes," *Catholic University Law Review* 64 (2015): 975–1006.

19. F. Schauer, "Fear, Risk and the First Amendment: Unraveling the Chilling Effect," *Boston Law Review* 58 (1978): 685–732.

20. T. Grundy, "Hong Kong's national security laws are designed to make the media self-censor" (2020), https://www.theguardian.com/world/2020/jul/14/hong-kongs-national-security-laws-are-designed-to-make-the-media-self-censor.

21. C. Wright, "Emily Self-Censorship and Associational Life in the Liberal Academy," *Society* 56 (2019): 538–549.

22. C. A. Hyde and B. J. Ruth, "Multicultural Content and Class Participation," *Journal of Social Work Education* 38 (2002): 241–256.

23. F. Rose, *The Tyranny of Silence* (Cato Institute, 2014).

24. S. Olsaretti, *Liberty, Desert and the Market* (Cambridge University Press, 2004).

25. B. Williams, *In the Beginning Was the Deed: Realism and Moralism in Political Argument* (Princeton University Press, 2008).

26. J. Wolff, "Freedom, Liberty and Property," *Critical Review* 11 (1997): 345–357.

27. N. Honeycutt, "Silence in the Classroom: The 2024 FIRE Faculty Survey Report. The Foundation for Individual Rights and Expression" (2024), https://www.thefire.org/research-learn/silence-classroom-2024-fire-faculty-survey-report.

28. R. P. Wolff, *In Defence of Anarchism* (Harper and Row, 1970).

29. P. Pettit, *Republicanism: A Theory of Freedom and Government* (Oxford University Press, 1997).

30. P. Pettit, "Enfranchising Silence: An Argument for Freedom of Speech," in *Freedom of Communication*, ed. T. Campbell and W. Sadurski (Dartmouth, 1994), 45–55.

31. S. Goldberg, *Conversational Pressure: Normativity in Speech Exchanges* (Oxford University Press, 2020).

32. R. Langton, "Disenfranchised Silence," in *Common Minds: Themes From the Philosophy of Philip Pettit*, ed. M. Smith, R. Goodin, F. Jackson, and G. Brennan (Oxford University Press, 2007), 199–214.

33. S. Shiffrin, *Speech Matters: On Lying, Morality and the Law* (Princeton University Press, 2014).

34. M. Friedman and J. Narveson, *Political Correctness: For and Against* (Rowman & Littlefield, 1995).

35. G. Loury, "Self-Censorship in Public Discourse," *Rationality and Society* 6 (1994): 428–461.

36. D. Moller, *Governing Least: A New England Libertarianism* (Oxford University Press, 2019).

37. A. Doyle, *Free Speech and Why It Matters* (Constable, 2021).

38. J. Narveson, "Political Correctness Revisited, Center for the Study of Ethics in Society Papers 42" (1998).

39. J. S. Mill, *On Liberty* (Batoche Books, 2001).

40. A. Meiklejohn, *Free Speech and Its Relation to Self-Government* (Harper & Brothers, 1948).

41. H. Joshi, "The Epistemic Significance of Social Pressure," *Canadian Journal of Philosophy* 52 (2022): 394–410.

42. C. T. Nguyen, "Echo Chambers and Epistemic Bubbles," *Episteme* 17 (2020): 141–161.

43. A. Piovarchy and S. Siskind, "Epistemic Health, Epistemic Immunity and Epistemic Innoculation," *Philosophical Studies* 180 (2023): 2329–2354.

44. H. Joshi, *Why It's OK to Speak Your Mind* (Routledge, 2021).

45. J. R. Otteson, "Escaping the Social Pull: Nonconformists and Self-Censorship," *Society* 56 (2019): 559–568.

46. R. M. Simpson, "Self-Censorship: The Chilling Effect and the Heating Effect," *Political Philosophy* 1 (2024): 345–380.

47. S. Moreau, *Faces of Inequality: A Theory of Wrongful Discrimination* (Oxford University Press, 2020).

48. S. Shiffrin, "Egalitarianism, Choice-Sensitivity, and Accommodation," in *Reason and Value: Themes From the Moral Philosophy of Joseph Raz*, ed. R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (Oxford University Press, 2004), 270–302.

49. C. Laborde, *Liberalism's Religion* (Harvard University Press, 2017).

50. M. Hannon, "Is There A Duty to Speak Your Mind?," *Social Epistemology* 38 (2024): 274–289.

51. R. Dworkin, "We Need a New Interpretation of Academic Freedom," in *The Future of Academic Freedom*, ed. L. Menand (University of Chicago Press, 1996), 187–198.