Ethical Theory and Moral Practice (2019) 22:657–671
https://doi.org/10.1007/s10677-019-10004-y

Hate Speech and Self-Restraint

Simon Thompson[1]

Published online: 28 May 2019

Abstract
In this article, my aim is to consider under what circumstances, and for what reasons, individuals may freely choose not to speak hatefully about others. Even if not threatened with legal sanction, why might they decide not to say something which they think they have good reason to say? My suggestion will be that there are various pro tanto reasons for individuals to restrain themselves from saying what they wanted to say. To be specific, I shall argue that such reasons fall into three analytically distinct categories, which I shall describe as normative codes of civility, ethics and morality. Thus each of these codes may provide different sorts of reasons for not engaging in hate speech. My hope is that the results of this investigation may usefully inform a strategy which aims to combat such speech without resource to the law.

Keywords Civility · Ethics · Hatred · Morality · Pro tanto reasons · Self-censorship · Self-restraint

1 Introductory Remarks

Many states have laws in place that criminalize hate speech. A number of rationales have been given for such laws. In some cases, there may be a pragmatic emphasis on the maintenance of public order. In others, a more principled basis for such regulation is provided. To take just one example, Jeremy Waldron suggests that hate speech should be criminalized in order to protect all citizens' equal standing (2012). At the same time, a number of criticisms have been made of such laws. For some, they are quite straightforwardly an illegitimate restriction on the right to freedom of expression. On their account, this right should include the freedom to speak hatefully about others. Even those who support such laws may do so with a degree of hesitancy, since they share some of the critics' concerns. They may worry, for example, that hate speech regulation has a chilling effect on legitimate speech or that it is unwise to give the state the power to censor.

This article begins from the thought that, if it is accepted that hate speech is wrong, but also acknowledged that the legal regulation of such speech has its costs, then it is worth considering how it might be possible to discourage hate speech without recourse to the law. If there was

✉ Simon Thompson
Simon.Thompson@uwe.ac.uk

[1] University of the West of England, Bristol, UK

some way of reducing the incidence of such speech without using criminal sanctions, then this would be doing something good, without risking the bad that such sanctions may cause. In order to think this idea through, a number of questions need to be addressed.

One question concerns agency. Who or what would be responsible for carrying out such activity? In this article, I shall assume that it is permissible (and may even be obligatory) for states to engage in activity of this kind, which may be justified in various ways. For example, Corey Brettschneider would regard it as the legitimate exercise of the 'expressive capacity' of the state. Talking specifically about hate speech, he argues:

> On my view, the state should simultaneously protect hateful viewpoints in its coercive capacity and criticize them in its expressive capacity. In this way the state can protect the right to express hateful viewpoints and, at the same time, defend the values of freedom and equality against discriminatory and racist challenges (2010: 1006).

It may be argued further that the state has a duty to promote the values on which it rests, and in particular to try to inculcate these values in its members (see e.g. Lægaard 2016: 128n2). For instance, if the state depends on its members exhibiting a certain level of civility, then it would be duty-bound to try to promote such civility in them.

Another question concerns method. How would the state go about trying to reduce the incidence of hate speech without resort to criminal sanction? A complete answer to this question could identify a wide variety of methods by which a state may seek to achieve this aim. This might, for instance, include the technique of 'nudging', which seeks to alter people's behaviour in a desired direction (Thaler and Sunstein 2008). In this article, however, I want to focus on just one aspect of this complex question. To be specific, my aim is to consider under what circumstances, and for what reasons, individuals may freely choose not to speak hatefully about others. Even if not threatened with legal sanction, why might people restrain themselves from saying something which they think they have good reason to say?[1] My hope is that an understanding of the different sorts of reasons which people might have for restraining themselves from speaking hatefully about others could usefully inform a strategy which aims to combat such speech without resource to the law. It will not be possible to describe such a strategy here, but in my concluding remarks I shall suggest what some of its principal elements are likely to be.

This might seem like a futile – or at best a foolish – endeavour. In the public spheres of contemporary societies, the expression of hatred seems to be the 'new normal'. In this case, there may seem to be little point in trying to understand why some people may sometimes not speak hatefully about others. Whilst I would fully understand this reaction to my proposal, I would want to turn it around and argue that it is precisely because we live in such times that it is necessary to explore all means possible for tackling this problem, including the possibility of self-restraint. I certainly would not claim that the investigation I intend to conduct here could provide a silver bullet. In popular parlance, 'haters gonna hate'. In particular, if someone speaks with the express intention of inciting hatred against others, it will be very difficult – although, I hope, not impossible – to persuade them to desist by telling them why some people may choose not to express their negative views of others. I would claim, however, that the results of my investigation may be able to help inform one small part of a strategy which a state

[1] It should be noted that social sanctions can sometimes do the same work as legal sanctions by imposing certain costs – such as ostracism or shaming – on those who speak hatefully about others. In this article, I exclude social and as well as legal sanctions from my analysis to the extent to which the former, like the latter, *cause* people to restrain themselves rather than giving them *reasons* to do so.

may employ in order to reduce the level, intensity and frequency of hate speech. And, to repeat, we need all the parts of such a strategy that we can find.

2 Range of Relevant Acts

I want to begin my argument by clearly identifying the range of acts in which I am interested. What do I mean by self-restraint? What sort of acts fall into this category? Conversely, what are not instances of self-restraint as I am characterizing it? What acts of non-expression fall outside the range with which I am concerned? I shall proceed first by exclusion, by describing two classes of acts which fall outside my remit.

First, I am not interested in acts of non-expression in which someone does not say something since they fear the consequences if they do speak. Where hate speech laws are in place, many people who have hateful sentiments will not express them since they fear the punishment that may follow if they do. I put acts of this kind aside since, *ex hypothesi*, I want to explore the reasons why individuals may refrain from hate speech even when such speech is not legally proscribed.[2] For my purposes in this article, therefore, I shall assume that a minimal censorship regime is in place in which only speech that risks causing imminent lawless action is regulated. The United States probably comes closest to such a regime. For instance, in the case of *Snyder v. Phelps* (2011), the Supreme Court determined that members of the Westboro Baptist Church were entitled to protest at the funeral of an American serviceman, since their speech, however offensive or outrageous, enjoyed protection under the First Amendment.[3]

Second, I am not interested in cases of non-expression in which someone does not say something because they have nothing to say. People do say some things, but they don't say infinitely more. This is nearly always because they have nothing that they want, need, wish or think they should say on a particular matter. In the current context, when someone does not speak hatefully about another group since they have nothing hateful they want to say, then their reasons for not speaking are not of interest to me.

I am now in a position to describe the two key characteristics of the acts of non-expression with which I am concerned. First, an individual has something that they want to say. To explain what this 'something' is like, I need to provide a definition of hate speech. There are more or less extensive ways of defining such speech, according to which wider or narrower ranges of expression would fall into this category. For the purposes of this article, and drawing on UK legislation in particular, I shall stipulate that hate speech involves: (1) the use of abusive, insulting or threatening language which (2) is intended or could be reasonably foreseen to stir up hatred against (3) a vulnerable minority to which certain characteristics are ascribed. I would not claim that this is an uncontroversial definition of hate speech. It may be argued, for instance, that (1) should be removed, since a speaker may intend to incite hatred using language which is not abusive, insulting or threatening. However, since my aim is to try to identify as wide as possible a range of reasons for not speaking hatefully, I want to keep the

[2] I should note that there may be cases in which a person does not speak hatefully because there is a hate speech law in place, but they do so not because they fear punishment but because they are impressed by the reasons that law gives for believing that hate speech is wrong. This sort of act of non-expression *does* fall into the class of cases in which I am interested here since it is the reason for the law rather than the consequences of breaking it that has made the individual desist.

[3] See https://supreme.justia.com/cases/federal/us/562/443/.

definition of hate speech itself as broad as I can at this stage of my argument. In the cases with which I am concerned, then, an individual wants to speak in the manner thus defined.

Second, this individual chooses not to say the thing that they want to say. To put it informally, they hold back, remain silent, bite their tongue, swallow their words, and so on. But they do not do so because they fear the consequences if they did speak. Rather, they choose not to speak because they accept the validity of some reason for self-restraint. It may seem as if this second characteristic presents us with a paradox. If a person chooses to restrain themselves, then it may appear to follow that they no longer want to say the thing that they wanted to say. Later on, I shall seek to resolve this apparent paradox by arguing that people can have a reason to speak hatefully and a reason not to speak hatefully at the same time.

3 Phenomenology of these Acts

In order to present a more detailed phenomenology of the acts of non-expression which fall into the set with which I am concerned, I want briefly to compare and contrast my account with two other analyses of these acts.[4] Later on, these analyses will also prove useful by pointing to some of the reasons which people might have for self-restraint.

First, I want to look at John Horton's analysis of 'self-censorship', which he places between 'straightforward censorship' on one side, and 'self-restraint' or 'self-control' on the other. There are two features of this analysis which I want to discuss. First, self-censorship is a deliberate or conscious decision by the self-censor (2011: 97), so that the self in question can regard themselves as the author of the act of censorship (2011: 98). To put it another way, acts of self-censorship are not a 'fearful response to threats'; if they were, then they would fall into the category of 'ordinary' or 'straightforward' censorship (2011: 98). Like me, then, Horton puts aside cases in which someone is silent because they fear the consequences of speaking.

Second, Horton nevertheless insists that self-censorship is experienced as 'censorious' (2011: 97). He uses the terms 'self-restraint' or 'self-control' to refer to cases in which the actor simply shows tact or discretion. But in cases of self-censorship, the actor believes 'it would not be unreasonable' (2011: 99) to say what is on their mind, and indeed they may have 'a feeling of resentment' about having to restraint themselves (2011: 99–100). But despite feeling resentful about it, this actor nevertheless chooses not to say what they think it would not be unreasonable to say.

Like Horton, I want to put aside those cases he calls 'straightforward censorship'. Since I want to consider how hate speech may be discouraged without recourse to the law, cases in which a person is forced not to speak by the presence and action of an external censor fall outside of my purview.

Unlike him, however, I wish to keep cases of self-restraint in play. Horton is not interested in these since they lack what he calls a 'genuinely censorious dimension' (2011: 97), and so do not count as instances of self-censorship. In other words, when we exercise self-restraint we are 'acting entirely out of our own volition' (2011: 99), and do not experience any kind of countervailing desire to speak. An example of this phenomenon in Horton's essay might be this: 'We might, for example, quite properly feel ashamed of what we feel or think, and therefore prefer to keep it to ourselves' (2011: 101). I do not want to put this class of acts of non-expression aside because, even if a person

[4] As far as I can tell, very little else has been written specifically about the idea of self-restraint, and related ideas such as self-censorship, from a philosophical perspective. Of tangential relevance, there is a literature in law on the nature of judicial self-restraint (e.g. Posner 2012), and the ancient Greek virtue of sophrosyne continues to be discussed and applied in various ways (e.g. Carr 2001).

feels no resentment about their decision not to speak, their reason for making this decision may nevertheless be of relevance to my analysis. To pick up on Horton's own example, an understanding of the reasons why someone might feel shame may be of use to a political strategy designed to reduce hate speech by extra-legal means. Thus, when I talk about self-restraint, I mean to include both of Horton's categories of self-censorship and self-restraint.

Cook and Heilmann offer an alternative account of self-censorship in which they distinguish between self-censorship 'by proxy' and 'by self-constraint'. Their analysis begins from a fundamental distinction between 'public' and 'private' self-censorship. In the first of these, a censor prevents the censee from saying something. This type of self-censorship corresponds to what Horton calls 'straightforward censorship', and I have already explained why I am putting this sort of case aside. It is the second type of self-censorship with which I am particularly concerned. In this case, censorship is not coercive since it occurs in the absence of an external censor (2013: 180, 189). Here the non-speaker wishes to say something, they are able to say it, but they choose not to say it.

Cook and Heilmann then make a further distinction between two types of private self-censorship which is of particular interest to me. I need to quote them at length:

> we distinguish between two ways in which private self-censorship can be established. First, there are undoubtedly many cases of private self-censorship where someone censors him or herself by taking a point of view external to his or her own private perspective: for example, an individual who is a member of an association may reflect on what the norms of the association imply for what he or she should express when participating in the association. We label such cases as private self-censorship *by proxy*. Second, there are cases in which an individual formulates his or her own conception of what it is permissible to express: for example, a person may develop a personal code where it is deemed impermissible to express obscene language or to speak about money in public company. We label such cases as private self-censorship by *self-constraint* (2013: 187).

I think that Cook and Heilmann's analysis is useful because it alerts us to the variety of reasons for which people may decide to censor themselves. Thus they put associational or social norms (2013: 187), work-based codes of conduct (2013: 188), and even the rules of a sect which is no longer extant (2013: 189) into the category of self-censorship by proxy. And they place the standpoints 'of common decency, maximal utility or deontological morality' (2013: 190), amongst other things, into the category of self-censorship by self-constraint. I shall refer to most of these examples in the analysis which follows.

Having said all of this, I think that Cook and Heilmann are wrong to try to draw a hard-and-fast distinction between these two categories of private self-censorship, since I do not believe that 'external' and 'private' or 'personal' viewpoints can be distinguished as clearly as they think. One way of making this point is to argue that there are no purely internal 'sources of values and principles' (2013: 187) of the kind needed to sustain their distinction. This is quite clear, I think, when we look at their examples of such regimes, which include codes of decency and morality. Morality is never a purely personal matter, and even apparently personal standards of decency are derived from multiple social sources which inform ideas about what counts as respectable conduct in particular contexts.[5] For this reason, whilst drawing on their account, I shall not give their distinction between private self-censorship by proxy and by self-constraint a formal role in my analysis.

[5] The extensive literature on the history and sociology of manners clearly demonstrates this point. The *locus classicus* is Norbert Elias's *The History of Manners*, the first volume of *The Civilizing Process* (Elias 1982).

4 Pro Tanto Reasons

Having got a good fix on the nature of self-restraint, I now want to turn to the reasons that people might have for choosing not to speak hatefully about others. Before looking at three specific sorts of normative code that may be able to provide reasons of the sort I am looking for, it will be useful to say something about the nature of these reasons in general.

To begin with, some reasons may be decisive, in the sense that there are able to completely overcome an individual's desire to engage in hate speech. If they could simply be persuaded that hate speech is wrong, then as a result they may no longer have any inclination to express themselves in this way. In this case, it might appear, self-restraint is no longer an issue: this individual, lacking any desire to speak hatefully, has no need to hold themselves back.

It is my contention, however, that in an important range of cases the reasons not to act on the desire to speak hatefully, although sufficiently strong to cause the speaker to restrain themselves, are not strong enough to entirely expunge that desire. As I said earlier on, this may sound paradoxical. If a person has a reason not to speak hatefully, then surely they will not want to speak in this way, and thus they will feel no need to restraint themselves. I want to suggest, to the contrary, that the experience of feeling competing reasons both to do something and not to do it is in fact quite common, and that it is an experience commonly found in the sorts of cases with which I am concerned.[6]

In order to understand what is going on in experiences of this kind, it may be useful to think in terms of 'pro tanto' reasons for action. If I have a reason of this kind to do something, then it is 'to that extent' right for me to do it. I may, however, also have a pro tanto reason *not* to do the very same thing. In Maria Alvarez's simple example: 'The fact that a joke is funny may be a reason to tell it; but the fact that it'll embarrass someone may be a reason against doing so. In that case, I have a *pro-tanto* reason to tell the joke and a different *pro-tanto* reason not to tell it' (Alvarez 2017). In this situation, it is necessary to assess the relative strengths of the competing pro tanto reasons in order to determine whether I have an 'all-things-considered' reason to tell the joke or not.

To see how this sort of analysis might be applied to the case of hate speech, let us consider Horton's example of a religious believer who has something important that they want to say concerning what they regard as the truth of their religion and the falsity of others' spiritual views. To say that this person wishes to proselytise and to do so with 'missionary zeal' (2011: 103) is to say that they have a pro tanto reason for speaking thus. But this person is also aware that those whom they address may experience their speech as mockery or ridicule. This gives the religious believer a pro tanto reason for *not* saying what they want to say. In this case, they may decide on balance to hold back, even when they think that they have good reason to speak, because they believe that there are weightier reasons for them not to do so. They have decided, in other words, that they have an all-things-considered reason not to speak. As Horton

[6] It may be noted that my account of self-restraint here bears a certain resemblance to the well-known paradox of toleration. To tolerate is to disapprove, but to choose not to act on that judgement. But if you think something is wrong, why shouldn't you act to stop that wrong? Similarly, in the case of self-restraint, as I understand it, the question is: if you think you have got a good reason to say something, why not say it? John Horton solves the paradox of toleration by arguing that acting to prevent the wrong may 'involve more harm than good' (1994: 11). In particular, it may involve 'restriction of people's freedom' (1994: 13). In a similar way, I solve what might be called the paradox of self-restraint by arguing that in some circumstances my reason for saying something is outweighed by a different good reason for not saying it.

puts it, 'at least an element of self-censorship may seem the most appropriate response to the expression of religious differences' (2011: 103).

Now, it is important to note that this is not an example of hate speech. Rather, the speaker wishes to say something with propositional content about religion, but they may hold back so that their audience does not feel ridiculed. I see no reason, however, why this analysis cannot be adapted as necessary to fit the cases under consideration here. Even when a speaker's express intention is to stir up hatred against others, it may still be possible to find a competing pro tanto reason capable of persuading them not to do so (although in this case the reason will have to be all the stronger if it is to outweigh the speaker's intention).

5 Intentionality and Foreseeability

This last point suggests something else about the nature of reasons for self-restraint: namely, that the sort of the reason which will best persuade speakers to restrain themselves will depend on whether the particular act of hate speech in question is intentional or not and whether it is reasonably foreseeable or not. Putting these two variables together gives us four categories.

5.1 Intentional and Reasonably Foreseeable

Into the first category go those acts in which the speaker (a) intended and (b) could reasonably foresee that their act would stir up hatred against its object. Arguably, this is the most straightforward of the four categories of speech act that I wish to distinguish. Here the speaker's explicit aim is to incite hatred against some particular group, and they have good reason to think that their act will have this desired effect.

There are, of course, innumerable instances of acts which fall into this class. For example, in the UK in 2017, a man named Nigel Pelham was convicted of publishing threatening written material with the intention of stirring up religious hatred. Two years earlier, he had posted various comments on his Facebook page, including: 'what this country needs is a bomb a mosque day' and 'we must burn mosques to the ground' (Pitt 2017).

Speech acts falling into this category are likely to be the toughest nuts to crack. If the speaker intends to stir up hatred, and if they know their act is likely to do so, then it will not be easy to find a countervailing reason which they could accept for deciding to restrain themselves. My hope is that it will not be impossible to do so.

5.2 Intentional but Futile

It would be somewhat odd to describe this second category as one in which the speaker (a) intended but (˜b) could not reasonably foresee that their act would stir up hatred against its object. If any sense can be made of this possibility, it would be that certain speech acts, although intended to incite hatred, fail to do so (and the speaker should have known that this would be the case). Hence it is necessary to replace 'could not reasonably have foreseen that their act would stir up hatred' with something like 'should have foreseen that their act would *not* stir up hatred'.

If it is understood in this way, then perhaps there are acts which fall into this category. Consider, for example, the case of the Westboro Baptist Church mentioned earlier. Two of the slogans on their placards read: 'God hates you' and 'Fag troops'. If it is the case (i) that these slogans were attempts to incite hatred against US soldiers, (ii) that there was no chance that

these attempts would succeed, and (iii) that the Church members should have known that (ii) was true, then these would be speech acts falling into this class.

Does my analysis need to include cases of this kind? On the one hand, if even futile attempts to stir up hatred could be effectively discouraged, then the incidence of hate speech would be reduced. On the other hand, since such attempts *are* futile, discouraging them need not be a priority for my analysis. Before moving on, it is worth noting that any reason which could persuade individuals to desist from intentional and foreseeable speech acts could also persuade them not to engage in intentional but futile acts either.

5.3 Not Intentional but Reasonably Foreseeable

Speech acts fall into the third category when the speaker (˜a) did not intend but (b) could reasonably foresee that their act would stir up hatred against its object. Thus these acts have two features: first, although the individual had some reason for speaking as they did, that reason was not to incite hatred; and, second, this individual should nevertheless have known that a likely effect of their speech would be to increase that level of hatred.

To get a grip on this sort of case, consider Anthony Bamber, who in 2008 published and circulated a pamphlet entitled *The Heroin Trade*, in which he blamed British Muslims for the vast majority of this trade. According to the judge in her remarks to the jury at his trial, 'Mr Bamber says that his intention was to publicise his campaign with no intention to stir up religious hatred. You must be sure that when he did the act of distribution, he intended to stir up religious hatred'. Probably because they could not be sure, the jury found Bamber not guilty. For the sake of the current argument, I shall assume that Bamber's protestations were sincere: his aim was to reveal the truth about a state of affairs, rather than to stir up hatred against the group that he held responsible for it.

The first question that I need to ask about cases falling into this third category is whether the idea of self-restraint can be applied to them. It might be thought that, if the speaker does not intend to speak hatefully, then it would not make sense to try to find a reason capable of making them restrain themselves. However, it remains the case that the individual in question did have a reason for speaking as they did. In some cases, as I suggested above, they may have wanted to reveal the truth about some situation. In other cases, they may have wanted to recommend that some course of action be taken. For example, it is at least conceivable that a speaker shares the Westboro Baptist Church's view that homosexual activity should not be tolerated, intends to persuade others of the merits of that view, but does not wish to stir up hatred against homosexuals themselves.

If this is right, then the investigation I am undertaking here should include speech acts falling into this category. It makes sense to look for reasons which may be able to persuade individuals to restrain themselves from speaking, even if the incitement of hatred is only an incidental effect of their speech. In these cases, relevant reasons would be those capable of persuading such individuals that the importance of informing their audience about some matter, or of recommending some course of action to them, is outweighed by the importance of avoiding the stirring up of hatred.

5.4 Not Intentional or Reasonably Foreseeable

The final category includes those acts in which the speaker (˜a) did not intend and (˜b) could not reasonably foresee that their act would stir up hatred against its object. They had no wish to

incite hatred against some particular group and nor did they have good reason to think that their act would have this unwanted effect.

To take another example from the UK, in 2006 Jack Straw – the Labour Home Secretary at the time – wrote a newspaper column in which he said that he had 'felt uncomfortable' talking to one of his Muslim constituents whilst she wore a full-face veil. After reflecting on this experience, he went on to explain, he had henceforth asked such women to remove their veils.[7] Once more for the sake of the argument, let us assume that Straw did not intend to stir up hatred and that he could not have reasonably foreseen that his article would have done so (see Brown 2008).

One might think that, in the absence of both intention and foreseeability, self-restraint can have no role to play. In cases such as these, the speaker would lack any awareness of the need to restraint themselves. In spite of this, I would argue that an analysis of the kind I am sketching here has relevance even in this sort of case. It is certainly true that, on the first occasion on which an individual's speech unintentionally and unforeseeably stirs up hatred, reasons for self-constraint would have no traction. If, however, the speaker is made aware of the effects of their speech after they have made it, then it may be that on future occasions they will accept the validity of certain reasons to restrain themselves.

Distinguishing these four categories of speech acts thus enables us to see more clearly what sort of reasons must be offered in each case to dissuade individuals from engaging in such acts. Where the incitement of hatred is intended, whether it is foreseeable or futile, then reasons for self-restraint must tackle hatred head on. Either a decisive or a pro tanto reason must be given for not speaking hatefully. Where incitement is not intended but is foreseeable, then it is necessary to offer reasons capable of persuading individuals that the importance of their speech is outweighed by its harmful consequences. Where incitement is not intended or foreseeable, the individual must be shown what the effects of their speech were, and then – as in the previous case – it is necessary to convince them that in the future the good achieved by such speech is less important than the bad of its side effects.

6 Three Normative Codes

Having considered the nature of reasons for self-restraint in general, I now want to consider the reasons why someone might not speak hatefully about others. There are, I think, many different reasons for not doing so, and these reasons vary in nature, scope and strength. In order to try to impose some order on my investigation, I shall suggest that such reasons fall into three analytically distinct categories, which I shall describe – perhaps in too formal a manner – as normative codes of civility, ethics and morality. I do not claim that these three codes are exhaustive. There may be other codes – and other more specific reasons – which might persuade people not to speak hatefully. Nor do I claim that these codes are mutually exclusive. Someone might not speak thus for reasons drawn from more than one of these codes, and it might not be clear which combination of codes these reasons were derived from. Nor finally do I claim that there is only one code of each type of relevance to my inquiry. There may, for instance, be a number of different moral codes which may give people reasons not to speak hatefully. My hope rather is that, by placing reasons for self-restraint into one of these three codes, it will be possible to give a clear account of the range of significantly different

[7] See https://www.theguardian.com/commentisfree/2006/oct/06/politics.uk.

sorts of reasons that people may accept for not engaging in hate speech. I shall now sketch each code in turn.

6.1 Codes of Civility

According to my account, a code of civility is conventional in character, comprising a set of (mostly) extra-legal norms which govern the interactions of individuals who live together in a particular society (Lægaard 2016: 127). From the perspective of civility, these individuals are not regarded as members of a shared ethical community, or as people who have moral duties to one another. Rather, they are seen as strangers who unavoidably come into contact with one another since they share the same social space.

In these circumstances, the role of a code of civility is, as Derek Edyvane suggests, to provide 'a kind of non-negotiable bedrock of decency in human affairs' (2016: 349). If such a code functions well, then the result, to put it negatively, will be 'the containment of conflict' (2016: 348). To put it more positively, civility makes possible social cooperation or 'living together peaceably with others' (Bejan 2011: 414).

Different analysts propose a variety of synonyms for civility, including courtesy, politeness, decency and respectful behaviour. Thus a civil person is, for example, courteous to others by moderating their behaviour towards them. It may be further suggested that, by behaving in this way, we take account of others' feelings, and try not to offend or upset them. By doing so, it may be argued, our peaceful co-existence with those others becomes more likely.

A number of the commentators to whom I have referred thus far find a place for civility in their analyses. Horton considers civility to be a notion which may provide reasons for self-restraint rather than self-censorship. In this case, being civil or courteous is a matter of demonstrating 'appropriate consideration' to another by 'showing discretion or being tactful' (2011: 99).[8] What Cook and Heilmann refer to as codes of 'personal' and 'public' decency (2013: 187–8) are very similar to what I am calling codes of civility. Codes of decency, they suggest, give people reason to refrain from acting in certain ways in order not to hurt others' feelings or to cause them offence (2013: 186, 193).

Finally, Lægaard suggests that civility provides a lens through which the so-called Danish cartoons controversy may be viewed. On this account, both critics and supporters of the cartoons' publication invoked, amongst other things, rival codes of civility. Thus:

> The main criticism of the publication from non-Muslim quarters in Denmark … focused on civility: … Here, the controversy was explicitly … understood … as a concern that all responsible citizens should be polite, should restrain themselves and should consider the feelings of others, including Muslims (2016: 129; and see 132).

By contrast, Lægaard suggests, those who invoked the right to freedom of expression in support of *Jyllands-Posten* argued that Danish Muslims' criticisms of the newspaper placed them 'outside the Danish consensus on norms of civility' (2016: 132–3).

The question, then, is whether a code of civility could provide a distinctive sort of reason not to engage in hate speech. Might someone who wishes to express hateful sentiments choose not to express them because they believe that it would be uncivil to do so? The brief account of

[8] A little later on in his article, however, Horton also seems to suggest that, in the case of religious differences, 'at least an element of self-censorship' may be appropriate in order to avoid causing others distress or offence (2011: 103).

civility that I have just presented suggests two sorts of reason why they might. First, it might be possible to persuade some individuals to desist by arguing that continuing civil peace requires compliance with a code of civility, where that code rules out at least some instances of hate speech. But this is a rather abstract argument, requiring individuals to be moved by an account of the function served by this particular social practice. Second, other individuals might be persuaded not to speak hatefully by arguing that such speech upsets, distresses, shocks and traumatizes those against whom it is directed. Clearly this is a more concrete argument, intended to engage directly with speakers' sympathies. However, whilst pointing out the gravity of the unintended effects of hate speech may persuade some speakers to desist, it seems less likely that it would persuade those who actually intend to incite hatred against a particular group.

One example of this phenomenon in practice might be the decision of the UK press not to reprint the Danish cartoons. According to a report in the *Press Gazette*, the *Sun* newspaper explained its decision like this: 'the cartoons are intended to insult Muslims and the *Sun* can see no justification for causing deliberate offence to our much-valued Muslim readers'.[9] If certain assumptions hold, then the *Sun*'s decision could be seen as a case of self-restraint for reasons derived from a code of civility since it is motivated by its desire not to offend.

Another example of self-restraint for reasons of civility would be a case in which an individual had in the past deliberately expressed hurtful views about a vulnerable minority group, but who has now come to regret doing so since they have come to understand the severity of the offence they had caused to members of that group. For instance, in 2018 the comedian Kevin Hart apologized for homophobic jokes he had made in the past. In a series of tweets, he said: 'I sincerely apologize to the LGBTQ community for my insensitive words from my past', adding later: 'I'm sorry that I hurt people'.[10] Presumably, Hart now believes he has reason for restraint himself from making such jokes in the future.

6.2 Codes of Ethics

On the account I am presenting here, a code of ethics, like one of civility, is a set of extra-legal norms which give us reason to treat others in certain ways. However, whilst the former code enables us to coordinate our actions with strangers, the latter governs our relationships with those whom we regard as fellow members of our ethical community.

To see what is distinctive about a code of ethics, as I define it, I need to explain what I mean by an ethical community. Drawing on the work of Axel Honneth, the idea is that the collective identity of a society is characterized not just in terms of a concrete set of conventional practices or a formal system of laws, but also by a distinctive set of values which are linked together to form a more or less coherent scheme. Each particular society, in other words, has a set of 'ethical goals and values' that comprise its 'cultural self-understanding' (1995: 122). Honneth refers to this aspect of a society as its 'intersubjectively shared value-horizon' (1995: 121) or 'value-system' (1995: 124).

I want to suggest, then, that a code of ethics encapsulates and expresses the identity of an ethical community, and that the shared values at its heart give the members of that community reasons for acting towards fellow members and others in ways consistent with that code.

[9] See https://www.pressgazette.co.uk/british-press-refuses-to-print-muhammad-cartoons/.
[10] See https://www.vulture.com/2019/01/kevin-hart-homophobic-tweets-apologies-ellen-degeneres.html.

None of the authors whom I have mentioned thus far talk explicitly about codes of ethics as possible sources of reasons for self-restraint. It is nevertheless possible to see brief allusions to such codes in a couple of their contributions to this debate. Cook and Heilmann refer occasionally to 'social norms', asking if such norms can 'act as censors' (2013: 180). They refer with greater frequency to the norms linked to particular associations, such as firms, suggesting that in the case of 'private self-censorship by proxy … public motivations, such as the norms of an association, provide the point of view from which an individual censors him or herself' (2013: 187).

Of course, associations such as firms are much smaller than the sort of ethical communities I have in mind. Lægaard gets much closer to my sort of community – and my sort of code – when he discusses a process which he describes as 'the nationalisation of liberal values' (2007). In this process, he suggests, a liberal value such as freedom of expression – which is assumed to be of universal validity – becomes transformed into a 'non-negotiable national value' (2016: 133), such as a "Danish value" (2016: 131, 134). In the terms I am using here, then, the complete set of Danish values constitutes an ethical code which, amongst other things, expresses the ethical identity of the Danish national community.

Could such a code of ethics provide reasons capable of persuading some people not to engage in hate speech? I think that there are two principal possibilities here. Some individuals may be motivated by a desire to support and help perpetuate their ethical community. If doing so is incompatible with at least some instances of hate speech, then this might provide sufficient reason for self-restraint. Other individuals may believe that the ethical code of their community includes values which in themselves provide reasons not to speak hatefully about others. They may think, in other words, that ethically appropriate behaviour requires them to desist from such speech.

By way of example, Section 266B of the Danish Criminal Code makes speech 'by which a group of people are threatened, insulted or degraded on account of their race, colour of skin, national or ethnic origin, religion, or sexual inclination' a criminal offence. Let us suppose that, instead of (or as well as) a law, not speaking in this way is also part of the ethical code of the Danish national community. In this case, it might be argued that the cartoons should not have been published – and should not be republished – because this would be to act in a manner contrary to Danish values. To generalize, a speaker may decide not speak hatefully about others since they accept that to do so would be, in Lægaard's word, 'un-Danish' (2016: 133).

A rather different example is provided by the case of the Dutch politician, Joram van Klaveren. Since becoming an MP in 2010, he had made highly anti-Islamic speeches, calling for the banning of the Koran and the closing down of mosques. However, whilst studying the Koran in 2018, he converted to Islam, and subsequently expressed regret for the views he had previously expressed.[11] If the values of Islam may be regarded as those of a particular religious community, then this is a case in which an individual now has reasons for self-restraint which are rooted in the values of a particular ethical community.

6.3 Codes of Morality

The third sort of code which may provide individuals with reasons not to say something hateful about others is moral in character. If a code of civility enables us to coordinate our actions with strangers, and a code of ethics governs our relationships with fellow members of

[11] See https://www.telegraph.co.uk/news/2019/02/05/dutch-former-anti-islam-mp-says-become-muslim/.

our ethical community, then a code of morality regulates our conduct towards all of those whom we regard as moral persons or rational agents.

At one end of the spectrum, a code of this kind may be universal in scope, intended to govern our relationships with all human beings. As we head towards the other end of the spectrum, moral codes become narrower in scope. Most frequently, such a code may apply to the set of individuals who are members of the same legal and political order. This does not mean that such a code collapses into one of ethics. A moral code does not apply to people in virtue their shared identity as part of a particular community, but rather in virtue of their subjection to a common authority, in this case as citizens of a particular state.

It is possible to find some references to moral codes in the work of those who have analysed notions of self-censorship and self-restraint. When discussing the Danish cartoons, Cook and Heilmann mention 'morality', alongside 'taste' and 'civility', as possible reasons for self-censorship (2013: 178). Later on, when discussing what they call 'private self-censorship by self-constraint', they suggest that this 'involves an individual taking a standpoint different from his or her first person standpoint on his or her expressions, such as a second person standpoint of common decency, maximal utility or deontological morality' (2013: 190). If the first of these three terms suggests a code of civility, the second two can be understood as codes of morality.

Lægaard makes several references to the notion of respect. For instance, he suggests that some opponents of the Danish cartoons' publication argued that self-restraint in this case would be a way of showing 'respect for others as equal members of society' (2016: 132; and see 134). However, it appears that the sort of respect that he has in mind is not associated with something like the acknowledgement of others' capacity for rational autonomy. Rather it is understood as an attitude of civility which is expressed in a desire not to 'unnecessarily hurt [others'] feelings' (2016: 132). Finally, Horton remarks at one point that, unlike censorship, self-censorship is not 'morally objectionable' because 'it would not appear to involve any … intrusion on [the self-censor's] autonomy' (2011: 100). This is a passing remark, however, and refers to the autonomy of the self-censor, rather than the autonomy of those about whom they decide not to speak.

Whilst there are only passing references to what I am calling a code of morality in these commentators' works, it is not difficult to imagine what a more fully worked out example of such a code would look like, and how it might provide distinctive reasons for self-restraint. To take one of Cook and Heilmann's examples, if I subscribe to a form of act utilitarianism, and believe that the speech act I am contemplating would undermine the goal of achieving the greatest happiness, then I have good reason not to engage in that act. Or, if I subscribe to a Kantian moral code, then I might think that I should refrain from hate speech since I believe that such speech is incompatible with my principle which holds that humanity should be treated as an end in itself.

In both of these cases, I would suggest, it is possible to see how a code of morality may provide a distinctive set of reasons for thinking that I should not say something which I am free to say. Such reasons are not rooted in a practical desire to maintain civil relations with the strangers with whom I come into contact, and nor are they based on the values I share with other members of my ethical community. Rather, reasons grounded in a code of morality suggest that I should restrain myself from speaking hatefully about others since to do so would be to violate my moral obligations to them.

Continuing with examples of repentance, consider the case of Maajid Nawaz. When he was a young student, Nawaz became a member of Hizb-ut-Tahrir, in which capacity he had

espoused the cause of radical Islamism. But a series of events prompted a dramatic change of view, and in 2017 he became co-founder of the counter-extremism think tank the Quillian foundation. Nawaz now describes himself as a supporter of 'liberal democratic, civil liberties and human rights values'.[12] If certain conditions hold, then it looks as if Nawaz now has reason not to speak in a certain way, where that reason is rooted in a particular moral code.

7 Concluding Remarks

In this article, I have considered at least some of the reasons why individuals may refrain from speaking hatefully about others. I looked in particular at acts of non-expression in which an individual decides not to say something that they want to say because they accept the validity of some reason for self-restraint. I then argued that in some cases such a reason may be sufficient to stop an individual from wanting to engage in hate speech, but that in other important cases pro tanto reasons will be the best available. I also suggested that it is important to know whether the speaker intends to incite hatred, and whether they can reasonably foresee that their act is likely to do so. This makes a significant difference to the sort of reason which may be effective in each particular case. Finally, I distinguished three normative codes – of civility, ethics and morality – and suggested that each may be a source of distinctive reasons for self-restraint.[13]

In this investigation, my guiding assumption has been that this account of reasons for self-restraint can be used to inform part of a strategy which the state could implement in order to try to reduce the intensity and frequency of hate speech by non-coercive means. Of course, I have not tried to describe or justify such a strategy here. In thinking about what it might look like, and how it might be justified, Philip Pettit could provide us with a good starting point. In his book on *Republicanism*, he asks 'what the republican state can do to facilitate the appearance and operation' of the civility on which it depends (1997: 251). The positive part of Pettit's answer is that the laws of the republic must be known to be 'legitimate interventions in civil life' (1997: 252). The negative part is that the state should make sure that it does not act in a way that undermines the supply of civility which 'the intangible hand' would otherwise produce. But he thinks that measures to actively foster and promote that supply – such as by means of education – are likely to be counter-productive (1997: 253).

However, since I am less sanguine than Pettit about the possibility that civility can be spontaneously produced and maintained, I would want to argue for more active measures. First, on the assumption that a minimal censorship regime is in place, I think that the state still has a role to play in using its expressive capacity to explain why it believes that hate speech is wrong, even when it is not illegal. Second, and against Pettit, I believe that education could play an invaluable role in inculcating and maintaining the values of civility, ethics and morality. The next task I would need to undertake would be to explain in detail how I think that these various values could usefully inform a state strategy which aims to discourage hate speech without recourse to the law.

[12] https://www.theguardian.com/politics/2015/aug/02/maajid-nawaz-how-a-former-islamist-became-david-camerons-anti-extremism-adviser.

[13] I might note that it is not too important for me to establish water-tight distinctions between these codes, so long as distinguishing between them gives a useful indication of the range of possible reasons for self-restraint.

References

Alvarez M (2017) Reasons for action: justification, motivation, explanation. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (winter 2017 edition). Available at: https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/. Accessed 24 May 2019

Bejan T (2011) 'The bond of civility': Roger Williams on toleration and its limits. History of European Ideas 37(4):409–420

Brettschneider C (2010) When the state speaks, what should it say? The dilemmas of freedom of expression and democratic persuasion. Perspect Polit 8(4):1005–1019

Brown, Alexander. 2008. The racial and religious hatred act 2006: a Millian response. Crit Rev Int Soc Pol Phil 11/1: 1–24, 1

Carr M (2001) Passionate deliberation: emotion, temperance, and the care ethic in clinical moral deliberation. Springer

Cook P, Heilmann C (2013) Two types of self-censorship: public and private. Political Studies 61(1):178–196

Edyvane D (2016) The passion for civility. Political Studies Review 15(3):344–354

Elias N (1982) *The History of Manners*, the first volume of *The Civilizing Process*. Random House, New York

Honneth A (1995) The struggle for recognition. Polity Press, Cambridge

Horton J (1994) Three (apparent) paradoxes of toleration. Synthesis Philosophica 9:7–20

Horton J (2011) Self-censorship. Res Publica 17(1):91–106

Lægaard S (2007) Liberal nationalism and the nationalisation of Liberal values. Nations and Nationalism 13(1): 37–55

Lægaard S (2016) The case of the Danish cartoons controversy: the paradox of civility. In: Göle N (ed) Islam and public controversy in Europe. Routledge, London

Pettit P (1997) Republicanism: a theory of freedom and government. Clarendon, Oxford

Pitt B (2017) Inciting hatred against Muslims—why we need a change in the law. Available at: https://medium.com/@pitt_bob/inciting-hatred-against-muslims-why-we-need-a-change-in-the-law-4cd7be1b0147. Accessed 24 May 2019

Posner R (2012) The rise and fall of judicial self-restraint. Calif Law Rev 100/3:519–556

Snyder v. Phelps, 562 U.S. 443 (2011). https://www.supremecourt.gov/opinions/10pdf/09-751.pdf. Accessed 24 May 2019

Thaler R, Sunstein C (2008) Nudge: improving decisions about health, wealth, and happiness. Penguin, Harmondsworth

Waldron J (2012) *The Harm in Hate Speech*. Harvard University Press, Boston