1

Conceptual Frontiers in the Understanding of Hate Speech

1.1 Introduction

No serious attempt to answer the question 'What is hate speech?' would be complete without an exploration of the outer limits of the concept(s). Is it hate speech to call someone 'fatty'? Is it hate speech to say 'Rapists should be hanged, but only the black ones'? Is it hate speech for a Black person to call a White person a 'honky' or 'white devil'? Is it hate speech to label someone a 'male chauvinist pig' or 'TERF' if doing so is righteous? Is it hate speech to use the slur 'poof' against a straight person? If so, who is the relevant victim? Is it hate speech to say 'homosexuality is disgusting'? Is it hate speech to declare 'Bisexuals do not exist'? Is it hate speech to assert 'Transwomen are not women' or 'Transwomen are men'? Is it hate speech to call someone a 'cisgender woman' if she identifies as simply a woman? Is it hate speech to perform blackface? If so, what about womanface? If Holocaust denial is hate speech, does the same apply to even bare denial of specific facts relating to the Holocaust? And, what about denial of genocide or other atrocity crimes in general? If defaming a group of people identified by their religion is hate speech, does the same apply to defamation of religion itself? Is using threatening words or behaviour with intent to cause people distress, alarm, or harassment based on their protected characteristic hate speech or hate crime? And, what about using threatening words or behaviour with intent to stir up hatred against people based on their protected characteristics?

These questions show how the frontiers of hate speech spread out in many directions and occur at different levels of analysis, all of which stand in need of exploration. This book explores several conceptual frontiers including grey area examples whose status as hate speech is marginal, uncertain, or subject to reasonable disagreement; differences between social media platform content policies on hate speech and hate speech laws; divergence between national and international hate speech laws; and the sometimes fuzzy boundary lines or penumbral zones where concepts of hate speech end and other concepts begin.

https://doi.org/10.1017/9781009357111.001 Published online by Cambridge University Press

In this book we shall not rehearse longstanding debates about the rights and wrongs of enacting legislation to make hate speech unlawful.[1] Nor for that matter do we get into the weeds of more recent arguments for and against legislation requiring social media platforms to remove unlawful hate speech.[2] Instead, we provide a new conceptual analysis of the term 'hate speech' with a view to clearing up not only misconceptions about the core meaning of the term but also ambiguities that emerge at the edges of its application – both of which have a tendency to hinder the former debates.

This chapter establishes the context of our project and defends its theoretical and practical importance. Section 1.2 outlines the basic conceptual framework employed in the book, including the distinction between two concepts of hate speech and our twin-track approach to analysing them. We also highlight some of the pay-offs that flow from this conceptual framework. Section 1.3 explains what we mean by 'grey areas of hate speech', including identifying three underlying reasons or explanations why certain phenomena might end up falling into these areas, namely moral, semantic, and conceptual. We also try to motivate the significance and value of working to clear up the grey areas. Finally, Section 1.4 introduces and attempts to respond to the sceptical challenge that says, because the term 'hate speech' is linked to conceptual ambiguities, misleading connotations, an explosion of applications, and politicisation, it would be better to dispense with both the term and its concepts. We critically examine five main ways of responding to this sceptical challenge: rehabilitation, downsizing, abandonment, replacement, and enhanced understanding. We defend the final response as being the most promising and the overarching goal of the book.

1.2 Conceptual Framework

As much as the book is ultimately concerned with the question 'What is hate speech?', the more direct focus will be on these related questions: 'What is paradigmatic hate speech?', 'What is grey area hate speech?', and 'What is not hate speech at all?'. We take it as read that answering these questions requires techniques and approaches from several academic disciplines including law, hate studies, sociolinguistics, philosophy, semiotics, communication studies, internet studies, politics, and international relations. More specifically, in this book we employ conceptual analysis, doctrinal analysis, linguistic analysis, critical analysis, and diachronic analysis. Overlaying all of this is the particular conceptual framework we

https://doi.org/10.1017/9781009357111.001 Published online by Cambridge University Press

employ, namely a distinction between two concepts of hate speech and a twin-track approach to analysing them. This section explains and defends that framework in more detail.

Before doing so, however, we first need to make a wider point about the academic study of hate speech. In this book we follow customary practice among academic scholars of hate speech by explicitly mentioning particular examples including citing specific words that most reasonable people would find deeply offensive and hateful if used in other contexts and with different intentions and purposes.[3] Reading such words may be confronting for all readers, but for some readers it could trigger upsetting reminders of personal experiences of hate speech. Some readers may simply prefer not to read two White academics repeating the n-word, for instance, no matter the circumstances. We want to warn readers about these possibilities and apologise in advance for any distress caused. At the same time, however, we wish to express our allyship with all victims of hate speech through a rigorous and open academic exploration of the subject. Sometimes it is necessary to see words on the page in order to gain a proper sense of just how uncomfortable but also ambiguous they can be.

Of course, some scholars make a conscious decision not to cite well-known or widely recognised examples of hate speech or instead to replace some of the letters with hyphens so as not to repeat the offending words in full (e.g. 'N----r', 'The J-ps are coming'). Matsuda, for example, defends the latter approach as follows: 'to prevent desensitization to harmful words'.[4] However, the primary subject of this book is grey area examples of hate speech that are not well known. The secondary purpose is to interrogate how organisations and institutions handle these specific examples. As such, we believe it would be virtually impossible to write intelligibly about grey area examples of hate speech without citing them. This book represents a good faith attempt to explore the frontiers of concepts of hate speech. Our hope is that the reader will recognise this intention as well as critically engaging with our arguments. We also hope that the reader will not become desensitised to our examples and will not copy or misuse the examples for hateful purposes.

In light of the above, throughout the book we rely on the conventional distinction between citing or mentioning slurs or hate speech in an academic context, on the one hand, and using slurs or hate speech, on the other hand.[5] It is also worth emphasising that academic purposes are not the only prima facie legitimate reasons for citing specific examples of hate speech. Other prima facie legitimate reasons might include everything from judges iterating examples of hate speech when discussing the facts

of legal cases; through to legislators mentioning especially notable examples of hate speech in parliamentary debates about whether to ban hate speech, what hate speech to ban, and how to draft sound legislation; to social media platforms citing examples of the sort of content they deem hate speech.[6]

1.2.1 The Ordinary and Legal Concepts of Hate Speech

Central to our analysis is the fundamental insight that there is not one but two concepts of hate speech: the ordinary concept and the legal concept.[7] As a rough preliminary characterisation, the ordinary concept is related to social norms, social rules, and social sanctions, such as social media platform content policies that disallow directly attacking people with racial slurs, for instance, whereas the legal concept is intimately bound up with legal norms, laws, and legal sanctions, such as criminal laws that ban incitement to hatred, discrimination, or violence on grounds of race, for instance. Meta (Facebook, Instagram) implicitly relies on this broad distinction within its operational structure and division of labour. Whilst its 'terms of use' and 'legal compliance' teams focus on unlawful hate speech within given jurisdictions, its global 'community standard', internal content policy, and content moderation teams concentrate on disallowed though lawful hate speech.[8] This distinction is also reflected in the mandate Facebook has given to its Oversight Board, namely to only deal with suspected cases of disallowed though lawful hate speech.[9] In a similar vein, in 2022 the Committee of Ministers of the Council of Europe distinguished between hate speech prohibited under criminal law, hate speech subject to civil or administrative law, and less grave forms of hate speech that call for alternative responses and measures, including but not limited to victim-support, education, and counterspeech.[10]

However, the above characterisation and illustrations only tell half the story. It is equally important to understand the genesis of the two concepts of hate speech and this has to do with semantics. The semantics of the term 'hate speech' are both non-compositional and polysemic in that its meaning is not divinable from the words out of which it is composed and it is a term with more than one meaning.[11] In its basic or original meaning, the term 'hate speech' relates to certain forms of unlawful speech, but it also conveys another, much wider meaning that captures lawful speech. To be more precise, during the late 1980s a group of legal academics and journalists began to use the neologism 'hate speech' to classify and make generalisations about a particular group of laws.[12] At first glance, it might

seem a stretch to interpret Mari Matsuda's coinage of the term 'hate speech' as her introducing an exploratory concept since the laws and legal phenomena she was studying were far from new and had previously been discussed using terms like 'group defamation' and 'hate propaganda'.[13] However, Matsuda's focus on people's lived experience of hate speech both as victims and seekers of justice for victims almost certainly had 'net positive exploratory utility'.[14] At the very least, because of the work of legal scholars like Matsuda, the term 'hate speech' has now been taken up by lawmakers and legal professionals themselves more explicitly. But, and this next part of the story is crucial, the new bit of terminology was subsequently adopted by various other speakers and has 'taken on a life of its own'.[15]

The splitting of the general notion into two concepts of hate speech partly supervenes on divergence in the sorts of speakers who use the term 'hate speech'. Today the term continues to be used by legal scholars, legal journalists, lawmakers, legal professionals, and law enforcement officers who typically have the legal concept in mind. But the term is also used by political scientists, linguists, sociologists, philosophers, mental health professionals, journalists, politicians, activists, social commentators, artists, writers, film and television producers, diversity in the workplace trainers, business leaders, social media platforms, and ordinary members of the public. Of course, in some instances these people may also have the legal concept in mind. But sometimes they have the ordinary concept in mind.

In fact, social media platforms have played a significant role in popularising the term 'hate speech' in the sense that their adoption of this term has been pivotal in turning it from a niche bit of legalese into a commonplace and non-technical phrase that regularly crops up in the media, popular culture, and public discourse more generally. That social media platforms like Facebook have public-facing community standards disallowing 'hate speech' has helped to bring this term to the public's attention. Especially in an online environment, it is now almost as hard to avoid the term 'hate speech' as it is to avoid hate speech itself.

Importantly, the basic division of labour between legal systems and social media platforms in tackling hate speech has also helped to cement some of the differing social functions or purposes played by the two concepts. The term 'hate speech' in its ordinary sense implies, among other things, the moral objectionableness and social unacceptability of certain speech, as in, speech worthy of moral condemnation and social sanction. By contrast, the term 'hate speech' in its legal sense connotes, among other things, that there is a pro tanto reason for legal institutions and

quasi-governmental agencies to prohibit or regulate certain speech, as in, speech that merits legal sanction. That said, some innovations in the legal field have challenged this basic division of labour, including the rise of internet regulations in Germany (NetzDG), France (Loi Avia), and more recently the United Kingdom (the Online Safety Act) and the European Union (the Digital Services Act), that create legal responsibilities on the part of social media platforms and other internet service providers to record and remove unlawful content.[16] However, even when platforms have legal responsibilities to record and remove unlawful hate speech, it is clear they will continue to operate their own content policies and remove disallowed though lawful hate speech.

We shall delve deeper into the distinction between the ordinary and legal concepts of hate speech in Chapter 5, including presenting a new theory of not only how they overlap and diverge but also how they relate to, or interact with, each other. However, it is perhaps worth making the point at this early stage that, when it comes to distinguishing the ordinary and legal concepts of hate speech, unfortunately we are not especially well served by the English language. The term 'hate speech' hides the fundamental ambiguity between the two different concepts and this occlusion has some interesting linguistic consequences. For example, researchers have recently uncovered a tendency among journalists to write stories about examples of speech in their local communities which they describe as 'not hate speech, but … '. They use this phrase partly because they recognise the speech in question is not unlawful but also partly because they see it as problematic and hate speechy.[17]

If a society or organisation were starting from scratch and creating a new vocabulary to capture the two concepts of hate speech, perhaps it would be wise to create two different terms or phrases. For example, it might propose the literal pairing 'lawful hate speech' and 'unlawful hate speech'. Then again, this might create uncertainty about which term to use if people want to refer to speech which is not currently but ought to be unlawful in their view. Perhaps a more eloquent pairing would be 'hate speech lite' and 'hate speech'. But this may convey the impression that the former is not as serious as the latter, which could call into question the legitimacy of social sanctions in the eyes of the public. Another option might be 'disallowed hate speech' and 'unlawful hate speech'. The former certainly speaks to the way in which certain forms of hate speech are disallowed by social media platforms. That said, the ordinary concept *hate speech* is not synonymous with the rules promulgated by such platforms, even though the latter significantly influence and shape the former. If the ordinary concept were

synonymous with the rules promulgated by platforms, then it would be confusing to say something like 'Facebook allows this speech even though it is hate speech'. When reading Facebook's community standard on 'hate speech', it should always be borne in mind that the term means 'hate speech of the sort disallowed by Facebook'. Not all forms of hate speech in the ordinary sense of the term are disallowed by Facebook. For these reasons, we proceed with 'ordinary hate speech' and 'unlawful hate speech'.

What is the pay-off from distinguishing two concepts of hate speech? First, to say there is not one but two concepts is more faithful to the myriad and varied uses of the term 'hate speech' in contemporary societies. Second, distinguishing the ordinary and legal concepts of hate speech provides a framework with which to mark the different social functions or purposes a society and institutions and organisations can reasonably expect these concepts to play. Third, drawing the distinction may be needed to capture people's lived experiences, the fact that they experience certain words as hate speech even though the words are not unlawful. This is especially important in the current internet epoch – following the Internet of content, the Internet of services, the Internet of people, and the Internet of things – which many users experience as the Internet of hate.[18] Fourth, drawing the distinction can mitigate against the problem that focusing only on legal understandings of hate speech creates a tendency to work towards a more precisely and narrowly defined concept ignoring the relevance or social utility of having a concept that is less precisely defined and broader in scope but still meaningful. Finally, drawing the distinction may be indispensable to improving the quality of not only academic and legal argument but also public debate more broadly about the nature of the problem of hate speech and the rights and wrongs of using different measures to tackle it. For example, there may be some interests, rights, principles, and values that are better suited as reasons for and against using legal norms, rules, and sanctions to combat hate speech; and others that are more fitting as reasons for and against using social norms, rules, and sanctions to deal with hate speech.

1.2.2 Twin-Track Approach

Building on the distinction between the ordinary and legal concepts of hate speech, we develop a twin-track approach to analysing the two concepts. Beginning with the ordinary concept, in the past fifteen to twenty years there has been an explosion in the number and range of applications of the term 'hate speech' and this has created a vast system of examples.

Members of this vast system can be, and often are, connected to prototypes or exemplars of hate speech in the sense that they show a complex pattern of overlapping and criss-crossing similarities or family resemblance with the exemplars.[19] As a result, competent language users of various kinds are able to use the term 'hate speech' in ways their interlocutors can understand even without being able to articulate a formal definition.[20]

By analogy, think of the way objects in the solar system are composed of materials traceable to the solar nebular in the early life of the system but have varying degrees of similarity in composition to each other. To push the astronomy analogy even further, exemplars of hate speech, like Earth as a prototype for planets in our solar system, stick in the minds of competent language users and provide a basis for comparison. This is one of the reasons why ordinary people have a hard time thinking that Pluto is not a planet, even though the scientific community has downgraded it to the category of dwarf plant.[21] Just as the term 'planet' has the ordinary meaning 'something like the Earth', even though it also has a far more formal and detailed scientific definition, so the term 'hate speech' has the ordinary meaning 'something like calling Black people "niggers", comparing Jews to rats, or denying the Holocaust', even though it also has a more precise and narrow legal definition within given jurisdictions. In Part I of this book we shall argue that different examples of hate speech in the ordinary (nonlegal) sense of the term share a family resemblance based on the distinguishing qualities of target, style, message, act, and effect even though they lack a singular common feature, akin to the way different planets in the ordinary (nonscientific) sense of the term share a family resemblance when taking into account mass, density, size, orbit, and gravitational influence, even though they lack a singular common feature.

Of course, we do not mean to downplay the fact that just as it is useful for astronomers to more precisely define the term 'planet' and limit its scope of application, so it is useful for legislators to lend greater precision to the term 'hate speech' and to restrict the range of unlawful hate speech. But when it comes to the ordinary concept *hate speech*, part of its usefulness may rest in its imprecision and wide application.[22] It allows people to air grievances but also to disagree. The ordinary concept is useful precisely because its looseness provides a space for communication: a shared vocabulary yet with polymorphic semantics. If people can at least come together to talk about the 'hate speech' they see around them, then at least they are talking the same language, even if they are thinking about different examples or disagree about what the examples have in common with exemplars. In that sense the ordinary concept *hate speech* is an

intercultural concept, namely the sort of concept that can serve as 'part of a contact language between groups'.[23]

This partly explains why there can be some shared understandings even between people who invoke the pejorative 'hate speech' to condemn certain things, on the one hand, and people who attack this condemnation, on the other.[24] For example, some people may denounce phrases like 'These women are sluts and whores' as hate speech because they see this language as seeking to control women's behaviour, but other people decry the very label 'hate speech' as itself a means by which mainstream culture seeks to control certain groups.[25] Religious speakers whose own words have been condemned for expressing hateful stereotypes or enacting forms of discrimination sometimes respond by using the label 'hate speech' to describe what they allege their critics are doing to them, namely expressing negative stereotypes or enacting forms of discrimination.[26] Here people are presupposing distinguishing qualities of the ordinary concept *hate speech* in the very act of rejecting or rearming that concept.

In short, we believe analysing the ordinary concept *hate speech* as a family resemblance concept with a wide scope of application not only lends meaning to the myriad and varied uses of the term but also better serves the range of social functions or purposes that a society can reasonably expect the ordinary concept to play. This includes everything from expressing moral condemnation and social disapproval towards certain forms of speech; through to voicing grievances about injustice and oppression; and to providing a vernacular for disagreement not dissimilar to other useful phrases in politics such as 'That's bullshit!' or 'That's propaganda!'.[27] In Chapter 5, we seek to uncover and distinguish the full spectrum of social purposes served by the ordinary and legal concepts of hate speech, respectively.

We also believe our analysis of the ordinary concept *hate speech* paves the way for a more balanced treatment of some of the most highly controversial and antagonistic areas of discourse today. As we shall try to show in Chapters 3 and 4, based on a family resemblance analysis, both the slurs 'tranny' and 'TERF' can be hate speech; both miscategorising people who identify as men as 'women' and miscategorising people who identify as just women as 'cisgender women' can be hate speech; both calling Black people 'niggers' and calling White people 'honkies' can be hate speech; and both stereotyping Mexicans as 'rapists' and stereotyping Trump voters as 'racists' can be hate speech.

In addition, this balanced treatment could help to mitigate what we call hate speech spirals or hatred feedback loops.[28] For example, when speakers say things like 'Transwomen are not women' or 'Transwomen

are men', this can provoke the use of slurs like 'TERF' and 'Feminazi'. In turn, some people make generalisations about trans people such as that if you accidentally misgender trans people they will lose their temper and yell at you or that if you intentionally question whether trans people really are the gender with which they identity, then they will send you death threats. As a result, other people will accuse speakers who express these sorts of generalisations as peddling transphobic stereotypes. In response, some speakers will call it bigotry and intolerance to seek to misconstrue legitimate concerns about the behaviour of some trans people in yelling at women and sending them death threats as transphobia. In turn, some trans people may argue that it is a microaggression to ignore the fact that they experience generalisations made about them as hate speech or to downplay or minimise real hate speech as merely voicing legitimate concerns about trans people. In our view, analysing the ordinary concept *hate speech* as a family resemblance concept with a wide scope of application helps to put a different light on this sort of dispute. Perhaps we have antagonistic groups of people drawing on various different forms of hate speech to attack each other whilst simultaneously disagreeing about what hate speech is, all because the concept lacks a singular common feature. However, it would be wrong to see the lack of a singular common feature as the root problem. Indeed, the mere fact that the groups are able to communicate at all may be a good thing. If they consciously use the term 'hate speech' in its ordinary sense to voice their grievances, this could be at worst a pressure-valve and at best a gateway to further deliberation, mutual understanding, and even compromise.

Turning to the legal concept *hate speech*, it is associated with a smaller cohort of competent language users than the ordinary concept. Unsurprisingly, this alters how the term 'hate speech' is understood among those with the relevant competency. For example, Facebook[29] and Twitter[30] define hate speech in the ordinary sense of the term as a 'direct attack' against people based on protected characteristics, but it is far less common for legal scholars, legislators, and courts to define hate speech using this broad terminology. Ironically, when legal scholars do define hate speech as 'speech attacks' this is often for the polemical purpose of depicting a legal interpretation of hate speech so broad as to pose a significant danger to free speech and ultimately with the aim of rejecting the case for banning hate speech.[31] Of course, both Facebook and Twitter supplement their general definitions of hate speech with long lists of rules disallowing specific forms of hate speech, but even these forms are typically more expansive than the speech prohibited by hate speech laws.

Focusing on legal scholars for a moment,[32] some limit themselves to defining a specific legal concept *hate speech* that relates to a particular body of law and legal jurisdiction, such as the concept *the stirring up hatred offences in England and Wales*. Others describe a cluster concept associated with a particular class of hate speech laws, such as the concept *incitement to hatred laws*. Yet others characterise an umbrella legal concept *hate speech* that implicates a range of different clusters of hate speech laws, including laws criminalising group libel, laws disallowing insult or denigration, laws banning incitement to hatred, discrimination, or violence, civil rights laws prohibited discriminatory harassment, human rights laws disallowing group vilification, and tort laws relating to intentional infliction of emotional distress or injuria. Because these characterisations differ in terms of both which body of law is the focus and the levels of generality at which they operate, they often wind up saying different, often contradictory things about what hate speech is.

Techniques of conceptual jurisprudence can be used to analyse the legal concept(s) *hate speech* by investigating the language and content of given bodies of law, even if the laws do not themselves contain the term 'hate speech'. This could work by identifying certain common elements of the laws in question. In order to identify a relevant body of law in the first place, the researcher could search through statutes and case law looking for words that are indicative of, contiguous with, or serviceable proxies for the legal concept – for example, 'protected characteristics', 'hatred', 'contempt', 'enmity', 'racial superiority', 'racist propaganda', 'xenophobia', 'antisemitism', 'Islamophobia', 'homophobia', 'group defamation', 'vilification', 'insult', 'negative stereotypes', 'stigmatisation', 'humiliation', 'degradation', 'emotional distress', 'violation of dignity', 'threats', 'cross burning', 'racist fighting words', 'discriminatory harassment', 'hostile environment discrimination', 'advocacy, encouragement, or incitement to hatred', 'incitement to discrimination', 'incitement to violence', 'incitement to genocide', 'Holocaust denial', 'genocide denial'.[33]

Since the legal concept *hate speech* is not a hermetically sealed object, it may also be feasible to analyse the concept by studying other bodies of discourse besides law itself. However, the challenge in this sort of approach lies in selecting a corpus that allows researchers to reliably separate the ordinary and legal concepts of hate speech. For example, if the corpus is simply any journalistic news story containing the term 'hate speech', then it would be difficult to say for certain that the discourse provides a reliable guide to the meaning of the legal concept given that the journalists concerned may be switching seamlessly between the ordinary and legal

concepts. But if the corpus is selected by the researcher using the more selective filter that the news story must be about relevant legislation or case law, then the researcher must rely on pre-theoretical assumptions about relevance, namely pre-theoretical assumptions about what makes something hate speech law as opposed to some other kind of law. These assumptions may reflect the researcher's own intuitions about the concept and create a confirmation bias in the results.

At any rate, the gold standard for conceptual jurisprudence at least in the analytic tradition is to try to clarify the meaning and extension of legal concepts by developing sets of necessary and sufficient conditions for appropriate application. These conditions might be supported in different ways. A condition might be descriptively grounded in relevant bits of legal text or legal practices. Also, it might be normatively justified by arguing the condition shows the concept in its best light. In this book we do not attempt to meet this gold standard and shall not provide a full and substantive definition of the legal concept(s) *hate speech* based on a complete set of necessary and sufficient conditions for its appropriate application. Nevertheless, in Chapter 5 we do outline several purposes that a society might reasonably expect the legal concept to play and contrast them with purposes associated with the ordinary concept. In addition, in Chapter 6 we provide a limited formal characterisation of the legal concept *hate speech* comprised of a single necessary condition, namely that, formally speaking, this concept only refers to laws which create bespoke crimes or other sorts of offences that do not have corresponding or parallel basic or base versions. This necessary condition for applying the legal concept *hate speech* is the mirror opposite of a well-established necessary condition for applying the concept *hate crime*.

The merits of this form of analysis include that it provides sufficient detail about the nature and extension of the legal concept *hate speech* as to establish a clear map of where this concept ends and where at least some other legal concepts begin without necessarily having to provide a full account of the disposition of this concept in relation to all concepts. Moreover, the analysis furnishes a characterisation of the legal concept *hate speech* that is still universal enough to be grounded and/or normatively justified in different ways, including being grounding in several different bodies of law and justified by the social functions or purposes a society can reasonably expect the legal concept to play. In addition to this, the analysis can be grounded in legal texts and legal practices at both the domestic and international levels. We attempt to demonstrate these merits in Part II of the book.

1.3 Grey Areas

Grey areas of hate speech are one of the main frontiers to be explored in the book. We use the term 'grey areas' in a broad way to mean examples that do not readily conform to classification. For example, examples can be hard to classify due to fundamental ambiguity at the conceptual level but also because of uncertainties based on limited information or reasonable disagreement about the examples. Grey area examples are not manifestly hate speech but could potentially be marginal hate speech or could turn out not to be hate speech pending further assessment. If paradigmatic examples of hate speech occupy space where the ordinary and legal concepts of hate speech overlap with each other, then marginal hate speech can be found at the outer edges of both concepts. These sorts of grey areas represent penumbral zones where concepts of hate speech overlap with other sorts of concepts. That said, an example might fall into a grey area simply because people do not know enough about the example or because the moral quality of the example is the subject of reasonable disagreement. In other words, cutting across different grey area examples is a further distinction between different underlying reasons or explanations for why certain types of words, signs, or behaviours might end up falling into grey areas of hate speech. We propose three main reasons: moral, semantic, and conceptual.

1.3.1 Moral Uncertainty and Disagreement

One reason for an example falling into a grey area of hate speech is moral uncertainty and disagreement. One of the many social purposes served by the term 'hate speech' in its ordinary sense is to express moral condemnation and social disapproval of certain forms of speech and by implication the speaker.[34] The term is in that sense a pejorative, but more than merely expressing disapproval, the term also conveys (expresses, implies, presupposes, infers) ideas about why certain things deserve disapproval. The ordinary concept *hate speech* is a thick normative concept combining description with moral evaluation. The concept says something not only about the basic semantic and/or pragmatic features of the speech to which it refers (e.g. related to protected characteristics, related to messages of hatred, contempt, and prejudice, related to the performance, justification, and perpetuation of subordination, misrecognition, and marginalisation) but also about its moral quality (e.g. wrong, unjust, bad,

https://doi.org/10.1017/9781009357111.001 Published online by Cambridge University Press

harmful, reprehensible, or otherwise morally objectionable), and the way moral quality relates to basic features (e.g. morally objectionable because related to protected characteristics, because related to hatred, contempt, and prejudice, and because related to subordination, misrecognition, and marginalisation).[35] However, if part of what is conveyed by the term 'hate speech' in its ordinary sense is a moral evaluation of the speech and the speaker, then this opens up the possibility of moral uncertainty and disagreement.

One source of moral uncertainty and disagreement has to do with ambiguity about the object of the moral evaluation. For instance, is the use of racial slurs morally objectionable due to the immoral ideas conveyed, due to the immoral sentiments expressed, due to the immoral act of communicating the ideas or expressing the sentiments to others, due to some kind of immoral act that is performed over and above the act of simply using the slur, due to its immoral effects, or due to a combination of some or all of these things? There may be uncertainty and disagreement over the status of a particular racial slur as hate speech because people have in mind different things about the slur that are morally objectionable.

In addition to this, moral uncertainty and disagreement could be rooted in fundamental differences of opinion about what the relevant immorality or moral objectionableness of hate speech consists in (e.g. bad intentions, violation of dignity, psychological damage, acts of subordination, misrecognition, and marginalisation, consequences for society including threats to public order or undermining democracy and legitimacy), the hierarchy between different forms of moral objectionableness, and the application of aspects of moral objectionableness to concrete circumstances. The upshot is that there could be dispute about whether certain words qualify as hate speech because people reasonably disagree about whether the words are morally objectionable all things considered. This is especially prominent in cases of what we call righteous attacks such as calling someone a 'male chauvinist pig' or 'TERF', to be discussed in Chapter 3. In one sense calling someone a 'TERF' might be a righteous thing to do but in another sense it might be harmful.

Furthermore, it goes without saying that whether speech is morally objectionable all things considered may be contextually dependent including both the specific conversational context and the wider social context. As a result, some forms of moral uncertainty and disagreement about the status of given bits of language as hate speech may reflect imperfect information or differences of interpretation about the relevant context as well as differences of opinion about how far context should in fact matter to moral evaluation.

1.3.2 Semantic Ambiguity

Another reason for an example falling into a grey area of hate speech has to do with semantic ambiguity (and, in the case of signs, symbols, or gestures, semiotic ambiguity); which is to say, the expressive, communicative, or performative phenomena purported to be hate speech may be such that they have more than one meaning or convey different content. For example, there might be ambiguity in the intended target or subject and ambiguity in what is being said about the target or subject. In these sorts of cases, the status of something as hate speech changes dramatically depending on what content is conveyed, and this can lead to significant uncertainty and disagreement among competent users of the term 'hate speech'.

In some instances, semantic ambiguity can result from the relevant words being ambiguous words such as homonyms or polysemes, where the words have multiple basic or literal meanings. Consider the following example.

(1) The Jews in my town always like to visit the bank at the weekend.

Here whether statement (1) is a candidate for being labelled 'hate speech' depends to a large extent (although not fully) on which sense of 'bank' is intended.

(1a) *The Jews in my town always like to visit the financial institution at the weekend.*
(1b) *The Jews in my town always like to visit the land near the river at the weekend.*

In other instances, semantic ambiguity can result from words or whole sentences being associated with different kinds of conveyed content. In this book we use the phrase 'conveyed content' in a deliberately broad way so that it might capture some or all of: literal meaning; conventional meaning; expressive meaning; implied meaning or implicature (including speaker implicature, conversational implicature, and sentence implicature); presuppositional content; inferred content; the speaker's intended meaning or encoded content; receiver understandings or decoded content; speaker and hearer negotiated meaning; use-conditional content; and contextually dependent content.[36] Consider another example.

(2) They are very hard working for black people.

The ambiguity of (2) lies in the fact that it conveys not only a literal meaning but also further conveyed content (implicature) that is not part of what is directly expressed.

(2a) *They are more hard working than other black people.*
(2b) *Black people tend not to be hard working all because they are black.*

Perhaps the semantic ambiguities in (1) and (2) are not especially difficult to resolve by looking at one or a combination of the words used, the speaker, the recipients, and the context including the conversational context, social practices and conventions, wider social and political circumstances, and cognitive environment including beliefs and assumptions. However, other examples may be trickier, as we shall now try to demonstrate.

First, consider phrases or statements containing words that are semantically ambiguous as between referring to a country and referring to its people. One such example has recently been reviewed by Facebook's Oversight Board.[37]

(3) Sout ta-yote [Burmese].

During the case, Facebook explained to the Board that, according to its interpretation, the Burmese word 'ta-yote' contained in (3) 'is perceived culturally and linguistically as an overlap of identities/meanings between China the country and the Chinese people'.[38] Facebook also explained that because in its view the user did not 'clearly indicate that the term refers to the country/government of China', it determined that 'the user is, at a minimum, referring to Chinese people'.[39] As a consequence, Facebook removed the post under its community standard on hate speech. However, the Board took a different view as follows.

> As the same word is used in Burmese to refer to a state and people from that state, context is key to understanding the intended meaning. A number of factors convinced the Board that the user was not targeting Chinese people, but the Chinese state.[40]

Second, consider compound sentences containing independent clauses that include references to countries but other independent clauses that refer to a people but no direct and explicit link is drawn between what is being said about the country and what is being said about the people.

(4) Think about their attitudes about eating whale meat—I hate Japan.
(5) I abhor what they are doing in Xinjiang—death to China!
(6) Israel is a fascist state—these people are scum.
(7) COVID-19 originated in China—they all deserve to get sick and die.

In the case of both (4) and (5), the second independent clause in the sentence contains an attack in the relevant sense (e.g. a vilificatory or threatening remark), but it is about a specific country not a people, whereas the

first independent clause contains a pronoun referring to a people but it is not explicit to whom it refers and it may not be an attack in the relevant sense. In the case of (6) and (7) the situation is reverse, the first independent clause relates to a specific country but may not be an attack in the relevant sense, whereas the second independent clause contains an attack in the relevant sense but it is not explicit to whom it refers. A strict approach to interpreting these compound sentences might be to insist that they cannot be labelled 'hate speech' because the attack and the pronoun do not occur together in the same independent clause and so the sentences remain ambiguous. A more nuanced interpretive approach might be to say there could be co-referencing exhibited in these sentences based on reading across between the independent clauses. On this approach, if any part of the sentence is referring to a people and any part of the sentence is an attack in the relevant sense, then the entire sentence can be classified as hate speech even if the reference and the attack occur in different clauses, provided that it is clear from the sentence, the speaker, and the context that the overall conveyed content does incorporate an attack on a people. Dispute about whether to classify these compound sentence as hate speech may well result from reasonable disagreement about which interpretive approach is correct.

Third, consider statements that directly attack a property, identity, or characteristic, but do not explicitly attack the people who possess that characteristic.

(8) Homosexuality is disgusting.
(9) Transgenderality is a fraud.

Focusing on the literal meanings of statements (8) and (9) might lead a society or organisation to conclude these statements are not hate speech, especially if they embrace a rule of thumb that says hate speech is a direct attack on people as opposed to properties, identities, or characteristics. However, looking at the non-literal content conveyed by these statements, such as the implicature or inference that homosexuals are disgusting or that transgender people are frauds, could lead a society or organisation to conclude these statements are hate speech. Here the dispute about appropriate classification rests on disagreement about which is the most relevant form of conveyed content. We shall investigate these sorts of cases more fully in Chapter 4.

Fourth, consider hybrid slurs which, like personal insults, are typically used to attack specific individuals, but, like prototypical hate speech, also seem to attack the groups to which targeted individuals belong.

(10) She is a fucking bitch.

Disputes about whether (10) counts as hate speech may reflect differences in how men and women typically understand the connotations of terms like 'bitch'. However, these differences are not simply a matter of the qualities or traits being attributed to the targets. In a 1978 participant experiment study asking 168 American college students their opinion on what they thought a person generally means when describing another person as a 'bitch', the researchers found that the top four traits most closely associated with the term 'bitch' by men were 'Critical', 'Stubborn', 'Inconsiderate', and 'Dominant', whereas the top four traits most closely associated with the term 'bitch' by women were 'Inconsiderate', 'Critical', 'Insincere', and 'Deceitful'.[41] This difference might predict that men are more likely to classify 'bitch' as hate speech than women, but in reality the opposite seems to be true. Perhaps the more significant difference is that men are more likely to interpret 'bitch' as merely a personal insult, whereas woman are more likely to interpret it as a group-based attack and, therefore, hate speech. We shall return to these issues in Chapter 3.

Finally, consider statements that attack a group but are ambiguous as to scope, as in, it is unclear if they attack all, most, or some of the group. Some statements might be hate speech in a straightforward way because they involve unqualified behavioural statements about the whole group and because the message conveyed is clearly defamatory or vilificatory. Other statements may be manifestly not speech because they involve qualified behavioural statements about only a small proportion of a group and because they appear to be grounded in historical facts, for instance. However, another class of statements fall into a grey area because, even though they involve qualified statements about members of a group, it is unclear what proportion of the group is being attacked and because they convey stereotypes that go beyond the mere statement of historical facts. Compare the following examples.

(11) All Russians are war criminals and they commit war crimes because they are Russians.
(12) Some Russian soldiers have committed war crimes in the past.
(13) More Russian soldiers are war criminals than is true of other nationalities.

In a recent Facebook Oversight Board case involving a user who had posted a poem about Russian soldiers,[42] Facebook provided a strong suggestion as to how it would classify statements such as (11) and (12). Whilst statements like (11) would be treated as Tier 1 hate speech by

virtue of being dehumanising attacks on people based on their nationality, statements like (12) would be deemed to be not hate speech in that sense.[43] This reflects the fact that Facebook would consider the group of people referred to by the noun-phrase 'some Russian soldiers' in (12) as a 'non-protected group' under Tier 1 of its community standard due to these people 'representing less than half of a group'.[44] We shall critically examine this rule in Chapter 3. However, this still leaves a question mark surrounding statement (13). The question is whether the noun-phrase 'Russian solders' is correctly interpreted as referring to all, most, or only some Russian soldiers, and this remains semantically ambiguous. After all, even if statement (13) is true, it is compatible with several different inferences.

(13a) *All Russian soldiers are war criminals.*
(13b) *The majority of Russian soldiers are war criminals.*
(13c) *Only a small minority of Russian soldiers are war criminals.*

If there is dispute about whether to classify statement (13) as hate speech, this might reflect disagreement about whether to focus only on the literal meaning of (13) or also to consider inferred content, disagreement about whether to look at any inference (13) could invite a receiver to make or only some inferences, and, ultimately, disagreement about whether (13) infers (13a), (13b), or (13c). We shall explore the status of selective attacks in Chapter 3.

In addition to this, dispute about whether (13) counts as hate speech might also be a function of disagreement about whether stereotypes can still be classified as hate speech if grounded in reliable statistics. Interestingly, Facebook's community standard on hate speech does not disallow stereotypes per se but does disallow 'harmful stereotypes' which it defines as 'dehumanising comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence'.[45] We shall scrutinise this sort of conceptual linkage between hate speech and harmfulness in various chapters of the book. For now, however, we simply point out that other social media platforms do not appear to make the qualification that only harmful stereotypes may count as hate speech. Twitter's hateful conduct policy, for example, refers instead to 'content that intends to degrade or reinforce *negative or harmful* stereotypes about a protected category'.[46] This divergence between Facebook and Twitter speaks to a third reason for examples falling into a grey area of hate speech, to which we now turn.

1.3.3 Conceptual Confusion and Contestedness

The third reason for an example falling into a grey area of hate speech relates to conceptual confusion and contestedness, namely lack of precision in, and contest over, the meaning of the term 'hate speech' itself. This can take different forms. For one thing, an example may appear to fall into a grey area due to a failure to disambiguate the ordinary and legal concepts of hate speech. Thus, certain people may deny that something counts as hate speech because they have in mind the legal concept, whereas other people may believe it is hate speech due to the fact that they are thinking about the ordinary concept.

Some social media platforms have not done enough to eliminate this confusion. Arguably, Facebook's community standard on hate speech is more complex, detailed, considered, coherent, responsive to feedback, evolving, transparent, reflective of its authoritative internal policy, and subject to independent oversight than any comparable public-facing hate speech code of any social media platform in existence globally. Ironically, because Facebook has put its head above the parapet it has often faced the heaviest criticism; and sometimes entirely justified criticism. As we shall argue in Chapters 3 and 4, Facebook's community standard is full of ambiguities, inconsistencies, and blind spots – which is perhaps inevitable given its complexity. However, there is one area of conceptual confusion created by the wording of the standard that is easily avoidable. Nowhere in the standard itself does Facebook clarify that the standard is about defining hate speech in the ordinary sense of the term as opposed to defining hate speech in the legal sense of the term, as in, the sort of unlawful hate speech recognised in national and international hate speech laws.

Furthermore, even without conflating the two concepts of hate speech, there can be contests among competent users over the extension of each concept. Some contests concern which particular styles of speech ought to count as hate speech. Consider the disagreement between certain Muslim states and the rest of the world as to whether defamation of religion is a form of unlawful hate speech.[47] Other contests relate to whether particular characteristics and groups are properly considered 'protected'. Consider disagreement among some legal scholars about whether sexist speech targeting women should be considered hate speech in the legal sense of the term.[48] Similar disagreements occur in relation to the ordinary concept *hate speech*, such as policy differences between Facebook, TikTok, and Twitter over whether, in relation to hate speech, age should be treated as

a primary or secondary protected characteristic and whether hate speech includes or excludes targeted misgendering.[49]

At a deeper level, there may be differing views over whether hate speech should be specified in objective or subjective ways. Some people might think it right to specify hate speech based not on the personal feelings or opinions of either the speaker or target but instead on widely accepted norms or commonly recognised standards concerning what is taboo. However, other people might insist such an approach is flawed for the reason that these norms and standards can be themselves discriminatory. For example, to live in a society in which there is a widely accepted norm that treats the word 'nigger' as hate speech but not the term 'fatty' may be experienced by some people as discrimination through misrecognition of an identity they affirm (*fat and proud*) and believe ought to be protected (*don't call be 'fatty'*). Kylie Weston-Scheuber makes a similar point about gendered slurs. 'Unlike words such as "nigger" or "Paki" which automatically trigger an adverse response in the public domain and are commonly recognised as racial slurs, words used to and about women are not recognised as gendered epithets.'[50] This differentiated taboo status cannot be easily justified by differences in the proportions of derogatory and non-derogatory (reclaimed) uses of the terms 'nigger' and 'bitch', for example. Both words can be, and often are, used with reclaimed meanings,[51] but it is not obviously the case that 'bitch' is disproportionately used with its reclaimed meaning; it very frequently retains is patriarchal, oppressive, and subordinating meanings.[52] We shall return to these arguments in Chapter 3.

A connected reason people may have to reject an over-reliance on widely accepted norms or commonly recognised standards is that such norms and standards can mask variations in the language different groups find acceptable or unacceptable and ignore the extent to which social taboos disproportionately reflect the views of powerful or dominant groups. Interestingly, in 1987 Preston and Stanley undertook a large-scale participant experiment asking 164 American college students to give their opinion on the question of what is the worst thing people could say about other people, and found that on average 'bitch' came out as the top answer to the question 'What is the worst thing a man can call a woman?' and to the question 'What is the worst thing a woman can call a woman?'.[53] However, breaking down the results by the gender of the respondent, 'the men were likely to say that the worst thing a man could say about a woman was "lesbian", whereas the women chose words such as "bitch" or "mean"'.[54] This suggests social taboos may be unreliable guides to what

https://doi.org/10.1017/9781009357111.001 Published online by Cambridge University Press

should count as hate speech. Similar observations apply to dictionary definitions of slurs, for example. The precise ways in which words are defined as slurs can reveal the attitudes of the majority culture – attitudes that may be implicitly biased against minority groups. For example, if dictionaries define the meaning of certain slurs as *derogatory* or *disparaging*, this could place emphasis or blame on the persons doing the disparaging. By contrast, if dictionaries define the meaning of certain slurs as *insulting* or *offensive*, this could implicitly place the emphasis or blame instead on the persons who are targeted. Ironically, to define certain words as insulting or offensive may signal ambivalence towards the targets and could even be tantamount to blaming the victim: that somehow the true source of 'the problem' lies in how certain groups of people react to words. Furthermore, this asymmetry in the definition of slurs may reveal an implicit racial bias if divergent patterns of dictionary definitions consistently track along racial lines, so that slurs against White people are defined as disparaging, but slurs against Black people as merely offensive.[55]

Based on all this, some people may gravitate back to specifying hate speech in subjective ways. But once again this is likely to be contested ground. A definition of hate speech can be subjective in stronger or weaker ways. A definition of hate speech is strongly subjective if it defines hate speech simply as whatever the targets of speech are disposed to call 'hate speech'.[56] Yet some people may reject that sort of definition, especially as a definition of hate speech in the legal sense of the term, on the grounds that it arbitrarily privileges the personal feelings or opinions of the targets over those of the speaker or of an impartial third party. It effectively gives victims a sort of discursive veto, namely the power to classify as hate speech any mode of expression or message they do not like. By contrast, a definition of hate speech is weakly subjective if it defines hate speech in terms of a set of objective criteria – such as language that carries a message of racial inferiority, is persecutorial, hateful, and degrading, or is intended to demean – but then gives priority to the perspective of the victim when it comes to understanding whether a given bit of language satisfies the relevant criteria.[57] But even here some people may refuse this definitional approach, especially as applied to the legal concept, on the grounds that it incorrectly assumes that members of a victim group will reliably and accurately know degrading speech, for instance, when they see it, and, moreover, will have uniform understanding.[58]

Yet another form of conceptual confusion and contestedness comes to the fore when hate speech is conceptualised not using a set of criteria but instead by affirming that it is a family resemblance concept. As explained

above, we believe the ordinary concept *hate speech* is a family resemblance concept and that its proper use is based on nuanced assessment of complex patterns of overlapping and criss-crossing similarities between examples. In Chapter 2, we identify a set of prototypes or exemplars of hate speech, and in Chapters 3 and 4 we assess the extent to which a range of grey area examples resemble these exemplars. However, it is possible that some uncertainty or dispute about whether to classify a given grey area example as hate speech may result if the example is similar or alike in some ways to exemplars of hate speech but also dissimilar or unalike in other ways.

Therefore, at the end of Chapter 2 we seek to mitigate this source of ambiguity by providing an account of the five main distinguishing qualities of hate speech in the ordinary sense of the term, namely target, style, message, act, and effect, and by drawing a distinction between partial and global resemblance. Whereas a particular example might have a partial resemblance if it shows at least some similarities with exemplars of hate speech, the global resemblance test involves an overall assessment of whether the range, extent, degree, and salience of the similarities between the example and exemplars of hate speech across at least four out of five of the main distinguishing qualities justifies classifying the example as hate speech on the basis that it resembles the exemplars of hate speech as closely as the exemplars resemble each other. Moreover, according to our normative conceptualisation of the ordinary concept *hate speech*, the global resemblance test must itself reflect the functions or purposes a society and its major organisations can reasonably expect this concept to play.

1.3.4 Why Clearing Up Grey Areas Matters

It is important to clear up grey areas of hate speech both in the ordinary and legal senses of the term for several reasons. First, clearing up grey areas can serve to clarify the general positions that different sides take on how strictly to interpret the word 'speech' in 'hate speech'. For example, some people argue the use of racial slurs on college campuses is paradigmatic hate speech that transgresses civility norms, norms of public discourse, or ideals of deliberative democracy.[59] Other people are more ambivalent about applying the label 'hate speech' to such cases and argue it is much harder to justify campus codes if the problem they are supposed to tackle is classified as transgressive speech. They argue it is more appropriate and powerful to classify the problem as transgressive behaviour or 'discriminatory harassment'.[60] Social media platforms also appear to take different

positions on the speech versus behaviour dilemma. It is not merely a superficial difference that Facebook has a community standard on 'hate speech', whereas Twitter has a policy on 'hateful conduct'. These labels provide clues to substantive differences in the definitions they each provide: Facebook's standard underscores dehumanising speech and imagery, generalisations that state inferiority, and harmful stereotypes,[61] whereas Twitter's policy highlights these things plus inciting others to spread fear, to harass, or to discriminate.[62] We believe some grey areas provide interesting test cases for these positions. Consider blackface, blackfishing, and womanface. They typically involve a combination of speech and behaviour. Clearing up whether these really are hate speech and, if so, what makes them hate speech, can involve taking a position on whether the salient features are dehumanising speech and harmful stereotypes, or some sort of immoral behaviour such as epistemic injustice or exploitation, or all of the above. We discuss this further in Chapter 4.

Second, clearing up grey areas can provide a response to the familiar slippery slope objection made against the label 'hate speech' itself and hate speech laws and codes: that if society accepts what seem to be not obviously unacceptable applications of the label 'hate speech' and not obviously unacceptable hate speech laws and codes at the current time, then it is stepping onto a slippery slope that will carry it towards a set of obviously unacceptable practices in the future.[63] Slippery slope objections have been made against both the ordinary and legal concepts of hate speech. For example, in spring of 2022 the Canadian Liberal MP Ya'ara Saks gave a speech during a parliamentary debate on the Emergencies Act in which she claimed that the catchphrase 'honk honk' used by participants in the Freedom Convoy was in fact a dog whistle commonly used by neo-Nazis as code for 'Heil Hitler'. Some critics took to social media to reject her analysis based on the following slippery slope objection.

> Where it started: 'Holocaust denial' is hate speech. Where it is now: 'Honk honk' is hate speech. Where it's going: 'I'm not voting Liberal' is hate speech.[64]

In a similar vein, speaking at an international conference on hate speech in Budapest in April 2006, the then US ambassador to the Organization for Security and Cooperation in Europe (OSCE), Julie Finley, raised the following slippery slope objection to hate speech laws.

> Efforts to restrict hate speech represent a clear and present danger to robust political debate. Once we start down the slippery slope, trying to define a nebulous term like 'hate speech,' we are heading for the potential for abuse.[65]

A common denominator to slippery slope objections is the assumption that certain terminus points are obviously unacceptable but once on the slippery slope it is hard or impossible to avoid ending up at those points. Grey areas of hate speech can be invoked in slippery slope objections as either terminus points or steps along the way. For example, some people might make the slippery slope objection that if today social media platforms classify as hate speech the phrase 'Death to Jews', then in the future they will have to classify as hate speech the hyperbolic phrase 'Death to Liberals'. This sort of hybrid attack is a grey area of hate speech because though the attack uses violent language and is group-based, the relevant group is identified by political beliefs and affiliations rather than the standard protected characteristics of race, ethnicity, nationality, religion, gender, sexual orientation, and so on. We believe clearing up whether hybrid attacks really are hate speech and, if so, what makes them hate speech, can provide an answer to this slippery slope objection. If 'Death to Liberals' turns out to be hate speech and for good reason, then the terminus point may not be as unacceptable as it first appears. Then again, if it turns out not to be hate speech and major institutions and organisations are able to reflect this in their hate speech laws and policies, then the alleged slippery slope is not so slippery after all. We shall say more about this grey area in Chapter 3.

Third, clearing up grey areas could help to defuse social tensions between groups over highly contentious language. Uncertainty about whether particular language is hate speech can lead groups to enter into a war of words that can escalate into violence. For example, members of a minority group may end up labelling language that attacks core aspects of their identity (e.g. atrocity denials) as 'hate speech'. The minority group might respond angrily to the criticism by using their own violent language or threats against the group to which the speakers belong. The minority group might fail to recognise its own language as hate speech because it regards itself as speaking in self-defence (e.g. righteous attacks). All of this can further inflame tensions and increase the risk of violence. However, using robust and credible conceptual analysis to clarify to both sides whether these grey areas of language really are hate speech might hold out a slim chance of creating common ground and deescalating the war of words. We shall discuss righteous attacks in Chapter 3 and denialist speech in Chapter 7.

Fourth, clearing up grey areas could be a way of answering complaints against institutions and organisations of the sort that can be corrosive of trust and legitimacy if left unanswered. For example, some people complain that the way hate speech laws and codes are drafted and enforced involves a double-standard, namely that they crack down on conventional

hate speech against minorities but turn a blind eye to reverse hate speech against White people and men, for example. Sometimes these complaints are made in good faith; other times they are more cynical. But either way, these complaints can be corrosive if left unanswered insofar as they undermine people's belief that the relevant institutions or organisations can be trusted to do the right thing and even the belief that the relevant laws or codes are legitimate. Clearing up uncertainty as to whether reverse attacks are actually hate speech, one way or the other, provides a response to these complaints that could mitigate these corrosive effects. On the one hand, if it can be shown that reverse attacks are not hate speech, then it draws the sting from the complaint. On the other hand, if it turns out that reverse attacks are hate speech, then it can provide an impetus for institutions and organisations to work harder to ensure there is no double standard in either the drafting or enforcement of the relevant laws and codes. We shall say more about this in Chapter 3.

Finally, clearing up grey areas is important not only to improve the quality of hate speech laws and codes but also the depth of public understanding, both of which are important for protecting human rights. If institutions and organisations get it wrong about particular grey areas of hate speech, they can end up not merely failing to punish or remove content that really is hate speech but also restricting language that is not hate speech. Along similar lines, if the general population gets it wrong about examples, this can lead to language not being countered even though it is hate speech or alternatively people coming under social pressure to remain silent even though they are not hate speakers. Cancel culture is a contemporary phenomenon often associated with college campuses and social media but it is simply the most recent incarnation of a much older phenomenon J. S. Mill called 'the tyranny of the majority'. To name and shame, censure, call out, revoke invitations to speak, or participate in a pile on against someone for hate speech can inflict real reputational cost. The fear of being cancelled and the self-censorship this fear creates can have as great a chilling effect on people's speech as oppressive laws and codes. Given the stakes are so high for misapplications of the label 'hate speech', clearing up grey areas is paramount.

1.4 What Use Is the Term 'Hate Speech' Anyway?

A final way of motivating the importance of this book and the project of conceptual analysis in which it is engaged is to try to answer the following sceptical question. What is so useful about the term 'hate speech'

https://doi.org/10.1017/9781009357111.001 Published online by Cambridge University Press

that cannot be achieved using different terminology? One answer to that question is that the term serves as a vehicle for expressing moral condemnation and social disapproval against certain forms of speech.[66] Then again, it might be countered that the job of expressing disapproval against speech is performed by many terms besides 'hate speech' – for example, 'impolite speech', 'ill-considered speech', 'transgressive speech', 'disrespectful speech', 'unjust speech', 'dangerous speech', 'harmful speech', 'irresponsible speech'. So, what is so special about the term 'hate speech'?

Part of the answer is surely that the term 'hate speech' seems to be particularly powerful in the present climate. People often reserve the term for forms of speech which are especially problematic or divisive within diverse societies, as in, societies made up of groups who are identified by, and identify with, protected characteristics. This perhaps explains why people engaged in hotly contested debates surrounding institutional racism, religious intolerance, sexual and transgender politics, and national identity and immigration, to name but a few examples, sometimes resort to accusing the other side of engaging in 'hate speech'.

However, this answer only invites additional challenges. For one thing, the term 'hate speech' is not merely ambiguous as between two concepts but can also be associated with the myth of hate, the misleading connotation that all examples of hate speech are intimately bound up in some way with thoughts, feelings, emotions, or sentiments of hate or hatred.[67] Reflection on competent usage of the term 'hate speech' shows that not everything appropriately labelled 'hate speech' is connected with mental states of hate or hatred. In a similar vein, looking at competent usage of the term 'microaggression' reveals that not everything appropriately labelled 'microaggression' involves actual aggression or aggressive behaviour (in fact very few microaggressions are actually aggressive). But insofar as terms like 'hate speech' and 'microaggression' can have misleading connotations, this raises a question as to their suitability as categories used in discussions about what we owe to each other and what should be allowed or disallowed in matters of interpersonal communication. It is also ironic that one of the ambiguities surrounding the term 'hate speech' is whether microaggression can count as hate speech or whether these are mutually exclusive concepts.

Furthermore, as we have repeatedly emphasised in this chapter, the term 'hate speech' has undergone a significant expansion in applications since the late 1980s. This expansion has caused some scholars to signal a

note of caution about its social utility. In the words of the political scientist Katharine Gelber: '[The term "hate speech"] is used in such a wide variety of contexts that the distinction between hate speech and other types of speech has been elided, and it is losing its traction as a specific claim that speech that can harm in a specific way and to a sufficient degree to warrant government regulation.'[68]

In addition to this, the term 'hate speech' has become highly politicised. For example, in disputes over immigration controls the pejorative 'hate speaker' has become a sort of trump card (excuse the pun). Moreover, even though many political actors use this pejorative on a principled basis, others use it for purely instrumentalist reasons. For them, it has become an opportunistic or go-to move in rhetorical combat – a move that is capable of packing a punch, cutting through the white noise, and being used either offensively or defensively depending on the need.[69] Of course, because no side has a monopoly on this pejorative, it is frequently used tit for tat, with both sides trading the accusation 'hate speaker!'[70] The fact that the proper meaning of terms like 'hate speech' and 'hate speaker' is not simply the subject of reasonable disagreement but also highly politicised has led some scholars to raise doubts over its social utility. As David Boromisza-Habashi observers, invoking a different meaning has become 'an "easy out" to speakers publicly accused of "hate speech"'.[71] If '[o]ne challenge for antiracist rhetoric is to identify evaluative terms for communicative conduct that can be used to persuasively portray racist or discriminatory talk as a norm violation', then at least in some countries this challenge cannot be easily met by the term 'hate speech' insofar as its semantic contestedness renders it unpersuasive.[72]

Given the conceptual ambiguities, misleading connotations, expansion of applications, and politicisation associated with the term 'hate speech', and in the light of the myriad functions or purposes a society might reasonably expect this general kind of term to fulfil, would it not be better to simply abandon the term 'hate speech' altogether? Rather than analyse it, why not bury it in the graveyard of useless or defunct concepts? In other words, assuming Matsuda had, in effect, introduced a new exploratory or experimental concept, has the time now come to declare the experiment a failure and to cut our losses?

We believe there are five main ways of responding to this sceptical challenge: rehabilitation, downsizing, abandonment, replacement, and enhanced understanding. In this final section of the chapter we critically examine each of these different responses, and having ruled out the first four, defend the final response as the most promising.

1.4.1 Rehabilitation

The first response to the sceptical challenge argues that the myth of hate is itself a misnomer and that in fact hate is essential to the meaning of the concept *hate speech*. The first step in this rehabilitationist project is to see hate as potentially connected with different kinds of mental states (e.g. beliefs, perceptions, prejudgments, evaluations, appraisals, episodic emotions, motivations, action tendencies, prescriptive judgments). The next step is to use the label 'hatred' or 'hate sentiments' to pick out a particular combination or set of these mental states as being essential to hate. For example, Robert Sternberg analyses the concept *hatred* in terms of (i) anger, (ii) contempt, and (iii) disgust.[73] Teresa Marques associates the concept *hate sentiments* with (i) negative appraisals of outgroup members as malevolent or malicious just by being members of that group, (ii) action tendencies that go from revenge, social exclusion, or attacks to the destruction of the target group, and (iii) motivation goals such as the desire to harm, humiliate, or even kill the target.[74] The final step is to claim that all proper applications of the concept *hate speech* refer to speech that is connected with one such set of mental states, whether hatred or hate sentiments.[75]

However, this rehabilitationist project has several flaws. For one thing, without drawing on a plausible account of how competent users actually employ the term 'hate speech' and of the social purposes the concept *hate speech* can be reasonably expected to play, the choice of one set of mental states over another to stand as the essential characteristics of hatred or hate connected with hate speech looks arbitrary.

Moreover, it is not obvious that all meaningful uses of the term 'hate speech' by competent language users refer to forms of speech that are in fact connected with one or other of these sets of mental states. For example, ordinary people might label statements as 'hate speech' simply due to them expressing stereotypes, generalisations, or false beliefs of a descriptive nature. Consider these statements.

(14) All African Americans are brawny.
(15) All Jews are money lenders.
(16) All Muslim men rape young girls.

Competent language users may be perfectly willing to label statements (14)–(16) as 'hate speech', and to do so properly and meaningfully, even if the sets of mental states labelled 'hatred' and 'hate sentiments' are not involved when the relevant speakers make these statements, and even if competent users do not assume that these mental states are involved.

In addition, the rehabilitationist project is partially eliminativist in that the terms 'hatred' and 'hate sentiments' are used as placeholders for more sophisticated mental phenomena – more sophisticated than the mental states associated with the term 'hate' in folk psychology. Looking carefully at the sets of mental states proffered by Sternberg and Marques respectively, neither of them include ordinary hate, as in, a state of intense dislike that most people associate with the term 'hate'. In the process of rehabilitating the concept *hate speech* it appears ordinary hate has been eliminated from the picture. If hate is essential to the meaning of the term 'hate speech', then it is not hate as most people understand it.

Finally, in order to sustain the dogma that every instance of hate speech involves hatred or hate sentiments one could, through stipulation, keep expanding the meanings of the terms 'hatred' or 'hate sentiments' to swallow up putative counter-examples to the dogma. For example, if competent language users would call something 'hate speech' that involves not anger, contempt, and disgust but instead ridicule and grievance, then one could simply redefine the term 'hatred' to mean *either anger, contempt, and disgust or ridicule and grievance*. Alternatively, if competent language users would call something 'hate speech' that involves not negative appraisals, action tendencies, and motivation goals but instead descriptive beliefs and prejudgments, then one could simply redefine the term 'hate sentiments' to mean *either negative appraisals, action tendencies, and motivation goals or descriptive beliefs and prejudgments*. Such a process could go on indefinitely, so as to deal with emerging counter-examples. But this appears to be an ad hoc approach without intrinsic conceptual merit. It does not posit a fixed essence that supports sound and reliable inferences to new examples.[76]

1.4.2 Downsizing

The second main response to the sceptical challenge attempts to relaunch the term 'hate speech' with a new slimline semantics. This downsizing strategy involves turning back the clock so that the term 'hate speech' reverts to its original legal meaning. This confronts the 'problems' of the systematic ambiguity and expansion of applications of the term 'hate speech' by advocating that henceforth it is used only in the legal sense. For example, Gelber advocates that society and its institutions should stick with the term 'hate speech' for the legal concept but use some other term(s) for what we call the ordinary concept.[77]

However, there are once again important weaknesses in this response. For one thing, when it comes to applications of the term 'hate speech',

once the genie is out of the bottle it cannot easily be put back in. Words and concepts are living things that are born, live, and die through the practices of people, societies, and organisations. Practices are difficult to shape and even harder to stop. Other things remaining equal, it is easier to coin a new piece of terminology than it is to kill it off. A neologism can appear overnight and its use grow rapidly, but its demise can take much longer. It is one thing for a neologist or conceptual inventor to inspire others with the words, 'And I call this "hate speech"'; it is quite another for someone to change or reverse linguistic practices with the slogan '"Hate speech" should only mean unlawful hate speech and no more'.

Part of the reason for this is that linguistic norms, like other norms, require sufficiently widespread adherence to become embedded and internalised. In the case of the term 'hate speech', downsizing the term would require buy-in not simply from ordinary language users but also from major internet companies like Meta, Alphabet, and ByteDance (the owners of Facebook, Instagram, YouTube, and TikTok respectively). However, these companies have invested heavily in the terminology 'hate speech' over many years or even decades in some cases. This terminology figures prominently in everything from the wording of their content policies; to the training they provide to their moderators around the world; through to the setting up of oversight boards; to public engagement with their users and the media; and on to significant collaboration with quasi-governmental agencies, civil society organisations, academics, and other stakeholders. To expect these companies to switch terminology seems unrealistic given the lost investment and costs of change management involved.

More importantly, so long as the term 'hate speech' and the ordinary concept to which it refers serve social purposes or functions, people and organisations are unlikely to commit to the downsizing strategy; and rightly so. We shall say more about the different social purposes served by the ordinary and legal concepts of hate speech in Chapter 5, but we make two additional points here simply about the social utility of the term 'hate speech' itself. One is that even if it were true that the term 'hate speech' has lost traction as a designation for the legal concept (or could lose traction) because of the explosion of people using it in the ordinary sense, this might be offset by the social utility of having a term that can refer not only to the legal concept but also to a category of speech that is broadly harmful or wrongful but not unlawful (the ordinary concept). In other words, the social utility of retaining the term 'hate speech' in all of its capacious glory ought to be assessed in the round, just as the social utility of sticking with terms like 'fraud' is evaluated taking into account the fact that

they can be ambiguous between ordinary and legal meanings. If a society or its institutions were to insist on downsizing every bit of terminology that has come to have both ordinary and legal meanings, there would be a vast downsizing project afoot. Furthermore, in the case of ambiguous terms, it is not clear why the legal meanings should be automatically privileged, rather than, say, saving the ordinary meanings and calling on lawmakers and legal professionals to simply come up with different words (legalese) for the things they justifiably need to legally restrict. Just because the term 'hate speech' originally had only a legal meaning, this does not mean the law has an automatic right to custody.

In addition, it is not obvious that the term 'hate speech' has lost traction as a specific claim about speech of the sort that can harm in such a way as to warrant legal restriction. The fact is that advocacy groups, lawmakers, courts, and international organisations from across the globe continue to debate, agree on obligations regarding, enact, repeal, revise, apply, and either widen or limit the scope of hate speech laws in countless ways, on the basis of myriad reasons, arguments, and principles, and typically in response to a body of commissioned reports, public consultations, and inputs from experts, most of which touch on the issue of harmfulness – and they manage to do all of this whilst still using the term 'hate speech'. We believe they can do this because they are also capable of making it clear both to each other and to the wider public that they have the legal concept in mind. Consider once again the Recommendation of the Committee of Ministers to member States on combating hate speech cited above, which explicitly and directly distinguishes between hate speech prohibited under criminal law, hate speech subject to civil or administrative law, and forms of lawful yet harmful hate speech that should be tackled in other ways.[78]

Of course, that institutions should be working towards improving their definitions of the legal concept(s) *hate speech*, such as by making hate speech laws more precisely and narrowly defined, is a point well taken.[79] To serve legal purposes, concepts should be capable of standing as public legal norms, as in, norms that are both justiciable through the work of lawmakers, law enforcement agencies, and courts and also understandable by those members of the public subject to them. But it is not clear that the explosion of popular discourse employing the term 'hate speech' makes it significantly harder for the legal concept(s) *hate speech* to play this role. The vast majority of hate speech laws do not actually contain the term 'hate speech'. So, the mere fact that this term is used not only by lawmakers and legal professionals but also by ordinary people and social media

platforms need not be an impediment to lawmakers and legal professionals doing their job. After all, ordinary people use the word 'fraud' all the time and with countless nonlegal meanings, but this does not prevent lawmakers and legal professionals from devising precise and narrow definitions of fraud offences. In fact, during the period that has seen an explosion in usage of the term 'hate speech', arguably hate speech laws have become more, not less, precise and narrow, largely due to the work of legislatures and courts. In England and Wales, the newer stirring up hatred offences contained in Part 3A of the Public Order Act 1986 are narrower than the older offences in Part 3; in Spain the Holocaust denial offence set out in Art. 607.2 of the Criminal Code was replaced with a narrower offence in Art. 510(1)(c); and in South Africa the hate speech offences established under s. 10 of the Promotion of Equality and Prevention of Unfair Discrimination Act 2000 (PEPUDA) have recently been declared unconstitutionally overbroad by the Constitutional Court in *Qwelane v South African Human Rights Commission and Another* [2021] ZACC 22 – to cite just three examples.

1.4.3 Abandonment

The third main response to the sceptical challenge argues that the challenges faced by the term 'hate speech' in terms of conceptual ambiguities, misleading connotations, expansion, and politicisation are overwhelming and clearly point in the direction of abandoning the term altogether even if there is no obvious replacement. This is not a matter of arguing for an alternative that is like the term 'hate speech' but minus its problems. Rather, it is a case of abandoning having language and concepts that do the sorts of things that the term 'hate speech' and its allied concepts do. The claim being that society simply does not need that sort of language or conceptual framework.

However, to argue successfully for straight abandonment would be difficult, especially if the acid test would involve showing that some people would be better off and nobody would be worse off without the term 'hate speech' *or something like it*. We believe the argument cannot be made. In Chapter 5 we shall set out the various valuable social functions played by the ordinary and legal concepts of hate speech. That is our primary reply to the abandonment project. Even so, we shall supplement that reply with some observations immediately. These observations are intended to show not only that some of the features of these concepts are typical of normative concepts as a broader category of concepts but also that taking the

abandonment response seriously could suggest abandoning all normative concepts (reductio ad absurdum).

As explained earlier in the chapter, the ordinary concept *hate speech* is a thick normative concept that combines description and normative evaluation. The term 'hate speech' is also a pejorative conveying a negative connotation about its objects. Interestingly, the concept *slur* and term 'slur' also share these qualities.[80] These shared qualities may partly explain why ordinary language users are pre-programmed to accepting slurs as paradigmatic hate speech; other reasons being cross-over in the targets of many slurs and hate speech in general (see Chapter 2). Like many terms that refer to thick normative concepts, the terms 'hate speech' and 'slur' are subject to dispute concerning the extensions of the concepts they represent. However, few people seriously argue that society should kill off the concept *slur* or that nobody would be worse off and some people would be better off if society abandoned the term 'slur'. Partly this is because the term 'slur' plays a useful role in picking out forms of speech that are typically derogatory and group-based. Surely a similar story can be told about the term 'hate speech'.

Another feature of the term 'hate speech' is that, at least in its ordinary meaning, it designates a family resemblance concept. But, once again, this feature is hardly unique among thick normative concepts. Various such concepts have been analysed as family resemblance concepts – for example, *objectification*,[81] *power*,[82] *exploitation*,[83] and *fascism*.[84] This in turn can mean that, as Ann Garry puts it, some '[f]amily resemblances can be much messier and more politically laden than Wittgenstein's own examples of games or numbers.'[85] The ordinary concept *hate speech* is no different. It has become a political football reflecting the values, goals, and ideologies of different interest groups. Some groups will see a family resemblance between given words and exemplars of hate speech that other groups may not see and *vice versa*. For example, Democrat voters think the term 'hate speaker' befits Trump due to his racist and xenophobic dog whistles, whilst Maga voters apply the same term to Trumphobic detractors but not to Trump himself.[86] Then again, abandoning the term 'hate speech' for these reasons would also seem to imply abandoning a slew of other terms including 'objectification', 'power', 'exploitation', and 'fascism'.

Now it is also true that the term 'hate speech' can convey different kinds of negative connotations depending on its occurrence alongside different bodies of collocates ('semantic prosody').[87] Within public discourses associated with the alt-right but also civil libertarianism and certain forms of feminism, the term 'hate speech' often occurs alongside words

like 'censorship', 'political correctness', 'merely offensive', 'words people dislike', 'dogma', 'victim culture', 'competitive victimhood', 'no platforming', 'snowflake generation', 'wokeness', 'closing down options', 'a form of tyranny', and 'yet another means of subordinating women'. Here the negative connotation attaches to the term 'hate speech' itself. Within public discourses associated with liberalism, progressivism, critical race theory, and LGBTQIA+ doctrine, by contrast, the term 'hate speech' typically occurs alongside words like 'racism', 'xenophobia', 'homophobia', 'transphobia', 'TERF', 'historically oppressed groups', 'prejudice', 'discrimination', 'climate of hatred', 'incitement', 'violence', and 'genocide'. Here the negative connotation attaches to the objects of the term 'hate speech' (i.e. things that are hate speech). But once again the term 'hate speech' does not seem unusual in this regard; many other terms in public life convey different connotations depending on linguistic environment. Consider the term 'discrimination' which has different connotations when it occurs alongside words like 'positive', 'affirmative', and 'reverse', compared to when it occurs alongside words such as 'racial', 'gender', and 'unfair'.

Nevertheless, given the politically laden nature of the ordinary concept *hate speech*, has the term 'hate speech' now become a liability? If both sides use it to attack each other, and attach different meanings and connotations to it, how can using it do any good or move disputes forward? However, our reply to this sceptical challenge is twofold. We have already touched on the first. It is that the term 'hate speech' is far from unique in being weaponised in this way. Such weaponisation is an occupational hazard for any thick normative concepts and family resemblance concepts used in political speech and public discourse. Recall the old adage that in politics 'one person's terrorist is another person's freedom fighter'. Likewise, the pejoratives 'bullshit' and 'propaganda' are traded back and forth between political parties. None of this means, however, it would be better not to have such terms. Indeed, it would be hard to think of an agreed list of neutral descriptive terms that political antagonists could agree on as a replacement lexicon. Furthermore, it is no coincidence that terms which are prevalent in political speech and public discourse are typically contested. Part of their social purpose is to provide expressive outlets for fundamental disagreements about what the other side has done, about what matters, and about what should happen in the future. They provide a language for non-violent disagreement if not always an avenue for actual compromise. Thus, the pejorative 'hate speaker' might be a fixture of political rhetoric not in spite of but because of its heterogeneous, capacious, and contested nature.

Second, it is important to see that the pejorative 'hate speaker' is not a personal insult akin to 'asshole' much less an evaluative similar to the thin normative concept *wrongdoer*. According to our normative conceptualisation, the pejorative 'hate speaker' is not a blank canvass onto which a speaker can paint almost whatever meaning he or she chooses and still be using it properly; not, that is, when it is being used at its best. Precisely because the ordinary concept *hate speech* is a family resemblance concept, a given application of the concept is only appropriate if it satisfies the global resemblance test meaning that the speech to which it refers must resemble exemplars as closely as they resemble each other. We believe it does not take much comparative reflection to see that calling detractors of Trump 'hate speakers' fails the global resemblance test. We shall also argue in Chapter 3 that the pejorative 'hate speaker' itself is insufficiently similar to exemplars of hate speech to be considered hate speech; if that were not the case, perhaps it might be curtains for the concept.

1.4.4 Replacement

The fourth main response to the sceptical challenge argues for replacing the term 'hate speech' with something else. For example, Dirk Kindermann accepts that the 'political, legal and epistemic importance' of the term 'hate speech' in 'public discourse, in linguistics, the humanities and social sciences, and in the legal domain' 'can hardly be overstated' but at the same time argues for its replacement with a term and concept that is better.[88] The replacement project is importantly different to an abandonment project. The former does not suggest that society and its institutions do not need a term like 'hate speech'. Instead, the replacement project must simply show that society would be better off with an alternative term. Then again, this is still a tall order. The test is partly whether the replacement term can perform all of the valuable social functions already performed by the term 'hate speech', the same or better, as well as serving other valuable social functions better than the term 'hate speech'. In addition, even if the replacement term is better in its own right, it must be worth the costs and risks of change. After all, sometimes the perfect is the enemy of the good.

We believe that no version of the replacement project can pass this twin test. Starting with the second part of the test, there is significant existing utility in the fact that people are familiar with and roughly understand the term 'hate speech'. Because the groundwork has already been laid, people can communicate using this term with a fair chance of mutual

comprehension even if there may be reasonable disagreement as well. Depending on which replacement term is chosen, to replace a known term with something else could mean restarting the process of comprehension-building and this potentially means the loss of valuable communication during the transition period. There is also an opportunity cost to replacing existing terminology: a potentially significant amount of attention, effort, and money is expended on making the switch that could be better spent on embedding the original term within society and supporting the work of institutions and organisations in combating the problematic phenomena to which the term refers. Risks of change include the possibility that having convinced society to jettison the original term, there is a lack of consensus over the replacement term and society ends up without a replacement, just a jumble of different words used by different people, in different ways, and without resemblance. The net result of this would be the absence of a single unifying point of linguistic contact between different groups in society. The worst-case scenario could be that, in the absence of this unifying discursive framework, the amount and severity of speech hitherto referred to as 'hate speech' grows. If there is no agreed label for the speech in question, speakers might be emboldened since there is no longer a banner around which a coalition of the concerned can gather to call out the speech and to coordinate a collective response. The collective response might include developing a coherent division of labour between social media platforms and legal institutions, for example. A precautionary principle might suggest that it is better to stick with the devil you know.

Turning to the second test, one of the benefits of the term 'hate speech' is precisely its ambiguity, namely that it can refer to both ordinary and legal concepts. Kindermann focuses on the functions a society can reasonably expect the legal concept *hate speech* to play and argues another concept (*discriminatory speech*) would do a better job.[89] Even if he were right about that (and we remain unpersuaded as we shall explain below), it ignores the key point that the term 'hate speech' also refers to an ordinary concept that serves a set of non-equivalent but no less valuable social functions (see Chapter 5). This cannot be ignored when deciding whether to replace the term 'hate speech'. The ordinary concept *hate speech* has taken on its own life beyond the law and is used by people to refer to a diverse range of expressive, communicative, and performative phenomena. Imprecision and breadth are strengths of the ordinary concept even if they would be weaknesses of the legal concept if it were similarly conceptualised. What is more, one of the strengths of the term 'hate speech' is that it can facilitate a certain type of useful read across between ordinary

and legal meanings. For example, there may be utility in social media platforms disallowing a category of speech that is broadly harmful or wrongful but not unlawful, and using the familiar term 'hate speech' to refer to that special category of speech. Doing so may increase the chances that users will have at least a broad understanding of roughly what sort of speech is at stake (publicity), an appreciation of why disallowing it may be justified (salience), and a basic disposition to follow the rules (compliance). These benefits could potentially outweigh any disbenefits of users failing to properly grasp some of the important differences between hate speech disallowed by social media platforms and hate speech that is unlawful.

However, the fundamental flaw in the replacementist project is the lack of a worthy replacement for the term 'hate speech'. There is no suitable replacement for the term 'hate speech' in the English language that is without similar or other sorts of problems. For example, the term 'hate speech' is itself only one term in a longer lineage of comparable terms including 'group libel' and 'hate propaganda'.[90] Yet these terms are much narrower than the term 'hate speech' and refer to specific styles including defamatory statements, statements inciting hatred or discrimination, and statements expressing racial superiority. Historically, they have tended not to refer to other styles such as slurs, stereotypes, and genocide denial, for example. This is one reason Matsuda coined the term 'hate speech'.

Are there any other alternatives? In the 1950s Gregory Allport invented the term 'antilocution' to describe the first stage in his scale of prejudice; a model that seeks to explain the social processes that lead up to genocides. Antilocution is speech against an out-group, especially by political figures, that is subsequently followed by 'avoidance', 'discrimination', 'physical attack', and 'extermination'.[91] Once again, however, this term only refers to certain styles, such as negative comments about an out-group that are not directly addressed to the subjects of those comments. Matsuda introduced the term 'hate speech' so as to capture another important part of the victim's story, namely the experience of being directly addressed or targeted with racist hate messages, threats, slurs, epithets, and disparagement, and the ways this direct targeting 'wounds' its victims.[92]

Around the time Matsuda introduced the term 'hate speech', several other scholars used the alternative 'assaultive speech' also to capture this idea of wounding and to underscore not merely the analogies but also the overlaps between physical and verbal assault.[93] This term may be especially apt when describing some of the personal damage done by words including inflicting emotional distress, knocking people's confidence or self-esteem, intimidating or shocking people into silence, and

causing people medium- to long-term emotional, psychological, and even physiologically harm. However, one of the strengths of the term 'hate speech' – one of the things that renders it so useful – is that it also refers to expressive phenomena where the metaphor of assault seems far less fitting such as in the case of statements that damage reputation, bring people into contempt, misrecognise identity, undermine civic dignity, rank people as inferior, or incite hatred or discrimination.

More recently, some scholars have used the term 'offensive public speech' to refer to people's experiences of racist speech directly addressed to them in public places and the silencing effect of such speech (i.e. the way fear of violence causes them not to 'talk back').[94] However, the term 'offensive public speech' may convey a misleading impression that what is distinctive or problematic about the speech in question is that people targeted find it offensive or are offended by it. This misleading impression, even if not one that is intended by the scholars concerned, risks minimising the grave nature of the speech and implicitly puts the blame onto victims for the fact of being offended, as though these issues could be easily resolved simply by greater effort on their part not to be offended.

Other alternatives avoid this misleading impression by describing the speech in question in such a way as to make the grave nature of the speech unmistakable and by implicitly conveying the idea that the fault lies with the speaker and society, not the victim. Examples include 'extreme speech',[95] 'dangerous speech',[96] and 'atrocity speech'.[97] However, these terms are poor replacements for the term 'hate speech'. On the one hand, the terms 'extreme speech' and 'dangerous speech' fail to convey the idea of speech targeting certain groups identified by protected characteristics (as opposed to any group of human beings), dissimilar to the way the term 'hate speech' is ordinarily used. They also potentially cover a category of speech that is far broader than is ordinarily captured with the term 'hate speech' insofar as they cover statements that espouse terrorist ideology or incite or glorify terrorism, and forms of child pornography. On the other hand, not all speech that is referred to using the term 'hate speech' is extreme or dangerous or related to atrocities in the specific sense of speech that incites violence or atrocities or significantly increases the risk that its audience will participate in violence or atrocities against members of another group.

Other terminology, such as 'bigoted speech',[98] 'prejudiced speech',[99] 'discriminatory speech',[100] and 'subordinating speech',[101] certainly does convey the idea of speech targeting certain groups identified by protected characteristics, similar to the term 'hate speech'. However, these terms also

have other weaknesses. For one thing, the term 'discriminatory speech' suggests speech that enacts, facilitates, perpetuates, legitimates, or incites discrimination, or speech that itself relies on, reproduces, or is a product of, discrimination, where this connection with discrimination is enough to warrant legal restriction. Although this term might be useful for speaking about the sort of speech that should be made unlawful, it misses another useful social function served by the term 'hate speech' in its ordinary sense, namely to speak about the sort of speech that is broadly harmful or wrongful but not discriminatory in a specific sense that might warrant legal restriction. Consider reverse attacks (e.g. 'honky', 'white devil') that do not enact or rely on discrimination against historically vulnerable groups. Or consider righteous attacks (e.g. 'male chauvinist pigs') that at least in one possible sense do not enact unfair discrimination. Part of the value of the term 'hate speech' lies in the fact that it discursively equips a society and its major organisations to condemn or even disallow such attacks without subjecting them to legal sanctions (see Chapter 3). Similarly, the term 'subordinating speech' might be understood in a more restrictive way to mean speech that has a special force to subordinate certain groups, that has this force because of a background of structural inequality and oppression (or 'systemic discrimination'[102]), and that involves subordination *qua* harm that is sufficiently grave to warrant legal restrictions. But once again, the term 'hate speech' can also imply these things with the added bonus of also referring to forms of speech that are broadly harmful or wrongful but not necessarily in the aforementioned way. Indeed, the term 'subordinating speech' might convey the idea that the speech in question is a mere symptom of the real problem (e.g. a background of structural inequality and oppression) and not a fundamental or intrinsic problem in its own right. The advantage of the term 'hate speech' is that it can imply or presuppose that the speech is itself a fundamental problem, and this makes it more versatile, enabling it to be applied to reverse attacks and righteous attacks, for instance. Of course, defenders of the term 'subordinating speech' might see this versatility as a weakness and insist that it is absurd to apply the label 'hate speech' to reverse attacks and righteous attacks (reductio ad absurdum). But in Chapter 3 we shall argue that this reaction would be hasty and that reverse attacks and righteous attacks share sufficient similarities with exemplars of hate speech across the distinguishing qualities of target, style, message, act, and effect to justify using the label 'hate speech', at least in its ordinary (nonlegal) sense.

Furthermore, terms like 'bigoted speech' and 'prejudiced speech' may harbour misleading connotations analogous to those of the term 'hate

speech'. Just as the term 'hate speech' may convey the misleading connotation that the relevant speech is necessarily connected with mental states of hate or hatred, so the terms 'bigoted speech' and 'prejudiced speech' may imply that the relevant speech is necessarily connected with mental states of bigotry and prejudice.[103] Now clearly some of the speech in question is connected with mental states of bigotry and prejudice, but not all the speech referred to using the original term 'hate speech' is connected with these mental states just as not all the speech referred to using the original term 'hate speech' is connected with mental states of hate or hatred.

In summary, it appears there is no suitable term in the English language that could provide not merely a replacement but a worthy replacement for 'hate speech', namely no term that means something like what 'hate speech' means, that can better serve the various social purposes the latter serves or else better serve other social purposes, that does not suffer the same or similar problems of conceptual ambiguity, misleading connotations, expansion, and politicisation, that does not have different but equally significant problems, and that is sufficiently better to justify the costs and risks associated with going through the process of replacement. So long as there is no superior alternative, all things considered, the remaining alternative is to stick with the term 'hate speech' but to improve conceptual understanding of it both among the general public and institutions and organisations.

1.4.5 Enhanced Understanding

The final response to the sceptical challenge (our preferred response) is to stick with the term 'hate speech' and to recognise rather than minimise or ignore its ambiguities, misleading connotations, expansion, and politicisation, but at the same time to shed greater light on its meanings so as to enhance understanding. This is by no means a perfect solution, but we believe it is the least bad of the five main responses we have explored.

We believe the goal of enhancing understanding is served by providing a rigorous, credible, and constructive analysis of the term 'hate speech'. This conceptual project should not be complacent or biased towards the status quo. The term 'hate speech' and the concepts it designates must earn their keep. Also, it must be clear how conceptualising the term 'hate speech' in a given way helps to serve the purposes a society can reasonably expect it to serve. Of course, to make an argument that certain terms and associated concepts have particular social purposes is also part of the analysis. In that sense, the overarching aim of our normative conceptualisation is to achieve coherence between meanings and purposes.

To give an example, we believe recognising that the term 'hate speech' is ambiguous between ordinary and legal concepts and showing how these concepts serve different social functions may in turn help to improve the quality of public discourse about hate speech. It might help lawmakers and legal professionals come to see that ordinary people interpret the term 'hate speech' in ways that reach beyond the scope of law, and appropriately so. It could also provide an impetus for the former to be even clearer about what they are doing and why, such as to explain to the public why they seek to define hate speech laws even more precisely and narrowly. Conversely, it could drive both ordinary people and major organisations like social media platforms to be more explicit about when they are using the term 'hate speech' in the nonlegal sense, and why they are doing so.

As well as giving stakeholders a nudge to clarify which of the two concepts of hate speech they are talking about, enhancing understanding could also have more practical benefits in terms of tackling the problem of hate speech. For one thing, it could help people to see that the law is not the only source of order in the field of interpersonal communication. The Internet has become a site of contestation, including both hate speech itself and debates over where to draw the line between acceptable and unacceptable speech. Understanding that lawmakers and legal professionals do not have a monopoly on the term 'hate speech' might be helpful to these debates. When a social media company publicises the fact that disallowed hate speech will be removed, made subject to reduced distribution, or amended with warning labels, it serves to highlight the impact of social norms, social rules, and social sanctions on our everyday lives. If relatively few people would consider going to the police with a complaint about hate speech they have experienced, perhaps they would be more willing to report cases to social media platforms or to trusted flagger organisations. Coming to understand that the term 'hate speech' does not simply refer to unlawful hate speech but also to broadly harmful or wrongful yet lawful hate speech might provide the information people need to come forward and report cases, a crucial step in tackling the problem.

Now it is one thing to say the ordinary concept *hate speech* serves a range of social functions; it is another thing to provide a conceptualisation of that concept which explains how it can best serve those social functions. If the ordinary concept *hate speech* were so utterly vague and unendingly capacious that it could be meaningfully applied to almost any speech whatsoever, then it would struggle to serve its purposes. It is likely competent users would vote with their feet and stop using it, perhaps jumping onto the next exploratory concept that looks more favourable.

We suggest that the ordinary concept *hate speech* is able to serve myriad social functions and that competent language users continue to buy into the concept (at least for now) precisely because it is a family resemblance concept. People implicitly understand the similarities between its exemplars. This makes a practical difference to how far the concept can be stretched. It means, for example, that competent language users can see the many important dissimilarities between the terms 'hate speaker' and 'hater', based on the differences in target, style, message, act, and effect of the speech in question.[104]

In short, we believe that enhanced understanding is the most promising response to the sceptical challenge. The remainder of the book is an attempt to fulfil that promise. In particular, we shall focus on examples that stress test our conceptual framework, namely grey areas of hate speech. The goal is to provide a system of conceptualisations that can shed light on, and wherever feasible resolve, the grey areas (as being hate speech, marginal hate speech, or not hate speech at all) and in a way that is consistent with the purposes that a society and its institutions and organisations can reasonably expect the term 'hate speech' to play.