BRILL
NIJHOFF

EUROPEAN CONVENTION ON HUMAN RIGHTS
LAW REVIEW XX (2022) 1–43

European
Convention on
Human Rights
Law Review
brill.com/eclr

Hate Speech: A Pragmatic Assessment of the European Court of Human Rights' Jurisprudence

Alessio Sardo | ORCID: 0000-0003-1015-2641
Senior Researcher, University of Genoa, Genoa, Italy
alessio.sardo@unige.it

Abstract

This paper aims to offer a fresh start for addressing several conundrums relating to hate speech. The method of research combines a conceptual analysis with a possible model for evaluating the European Court of Human Rights' (ECtHR) decisions on hate speech. *First,* drawing on a Gricean account of communication, the argument proposes a working definition of hate speech: hate speech is best understood as a public speech act, aimed at subordinating individuals, which causes harm to targeted groups. *Second,* the paper offers a taxonomy of the different forms of hate speech, based on their degree of explicitness and detachment from the speaker's intentions. The most explicit forms of (harmful) hate speech – e.g., racial slurs, fighting words, or overtly sexist remarks – will be distinguished from implicit forms of (harmful) hate speech – e.g., innuendo, insinuation, and irony. *Third,* the author develops a categorical framework for hate speech that can be used as a standard for evaluating the jurisprudence of the ECtHR. The author also discusses three limitations of the model: a) the absence of a European consensus, b) puzzled speakers, and c) difficulty in determining harm.

Keywords

hate speech – Article 10 ECHR – pragmatics – Paul Grice – reasonable speaker's test

1 Introduction

John C Roberts, Chief Justice of the US Supreme Court, once wrote: 'Speech is powerful. It can stir people to action, move them to tears of both joy and

 | DOI:10.1163/26663236-BJA10054

sorrow [...].'[1] Similarly, the Proverbs (15:1) reads: 'The tongue has no bones, but it is strong enough to break a heart. So be careful with your words.' These considerations gain particular weight when applied to hate speech. Respectful and inclusive use of language is a tool for promoting tolerance and equality.[2] Hate speech, instead, might bring about social fragmentation and discrimination.[3] Derogatory speech is often seen as a threat to human rights;[4] this extreme use of language is usually associated with social conflict.[5] It does not necessarily follow from these premises that regulation is the best response, unless we advocate a 'militant' view of democracy (i.e., the notion that democratic orders shall curb the rise of extremist and anti-democratic discourses in order to preserve democracy itself).[6] Conceptually, democracies can be conceived in other ways. Eric Heinze, for example, argues in favour of unrestricted freedom of expression on the grounds that it is important to fulfil the 'legitimizing expressive conditions' of a 'longstanding, stable and prosperous democracy.'[7] In his view, banning certain categories of speech undermines the democratic process.[8] Moreover, imposing linguistic politeness does not necessarily prevent discrimination and violence.[9] However, the European Court of Human Rights (ECtHR, Strasbourg Court) has generally favoured a militant approach and, in particular, has grappled with hate speech in high-profile decisions, such as *Féret v Belgium*, holding that some forms of hate speech are not protected by

1 *Snyder v Phelps*, 562 US 443 (2011).

2 UK Preuß, 'Die empfindsame Demokratie', in *Verbot der NPD oder mit Rechtsradikalen leben?*, C Leggewie and H Meier (eds), (Suhrkamp 2002) 115.

3 S Assimakopoulos, FH Baider, and S Millar (eds), *Online Hate Speech in the European Union: A Discourse-Analytic Perspective* (Springer 2017).

4 *Erbakan v Turkey* 59405/00 (ECtHR, 6 July 2006) para 56.

5 See, generally, M Crock and L Benson (eds), *Protecting Migrant Children* (Edward Elgar 2018).

6 The concept of 'militant democracy' was coined by Karl Lowenstein (see, K Loewenstein, 'Militant Democracy and Fundamental Rights I' (1937) 31(3) *American Political Science Review* 638). Militant democracy denotes the necessity of curbing subversive, extreme, and, ultimately, anti-democratic movements as a self-preservation tool for preserving democracy. The historical practice of implementing this idea through a strategic definition of political participation is complex and tangled. For a discussion, see, G Capoccia, 'Militant Democracy: The Institutional Bases of Democratic Self-Preservation' (2013) 9 *Annual Review of Law and Social Science* 207; S Tyulkina, *Militant Democracy. Undemocratic Political Parties and Beyond* (Routledge 2015). For a normative analysis, see, for instance, AS Kirshner, *A Theory of Militant Democracy* (Yale University Press 2014).

7 E Heinze, *Hate Speech and Democratic Citizenship* (Oxford University Press 2016) 95, 145, and 210 and the following.

8 Ibid 86–87.

9 Ibid 150.

Articles 10 and 17 of the European Convention on Human Rights (ECHR).[10] Determining which expressions count as hate speech is, therefore, central for both understanding and evaluating the jurisprudence of the ECtHR.

The content-based restrictions on hate speech endorsed by the ECtHR deserve close attention. The Strasbourg Court has justified bans on both *explicit* forms of hate speech and more subtle, *implicit* (subtextual) forms of hate speech,[11] that the speaker can retract or cancel, invoking *plausible deniability*.[12] The ECtHR's strategy could respond to concerns about the potential danger of implicit hate speech. Refraining from overtly committing to the hatred content is a communicative manoeuvre that speakers use to avoid accountability. The use of implicit communication is consistent with a dissemination strategy: the transmission of hate through more subtle forms of backdoor communication. 'Anna is blonde but she is also smart. I am just saying' can be a sexist remark, although not entirely explicit. By partially hiding the implicit content of an utterance, the speaker can promote its effective dissemination through the audience's unreflective acceptance.

This paper defends a *pragmatic analysis* of hate speech, with an eye to implicit hate speech. Specifically, the analysis aspires to fill a gap in the existing literature by putting forward the following novel key arguments: *first,* to combine the *reasonable speaker test* – traditionally endorsed for reviewing First Amendment cases in the US – with a Gricean account of communication;[13] *second,* to use this innovative, pragmatic version of the reasonable speaker test for assessing the ECtHR jurisprudence on hate speech. Although this proposal could strike the readers as provocative at first glance, my argument will show that the Gricean version of the reasonable speaker test is fundamental for assessing whether a piece of communication *qualifies as* hate speech before deciding whether it can be restricted by Articles 10 and 17 ECHR. This newly-introduced pragmatic test may also be a response to critics of content-based bans on hate speech; by adopting the perspective of the reasonable, ordinary speaker, the ECtHR would not impose technocratic and legalistic definitions of hate speech, as judges are more likely to favour a citizen-centred and consensus-based perspective in determining whether an utterance constitutes hate

10 *Féret v Belgium* 15615/07 (ECtHR, 16 July 2009).

11 M Ignatieff, 'Respect and the Rules of the Road', in *Free Expression is No Offence,* L Appignanesi (ed), (Penguin 2005) 127, 128.

12 R Delgado and J Stefancic, *Understanding Words that Wound* (Routledge 2019) 11.

13 For a model based, instead, on the 'reasonable target', see, PY Kuhn, 'Reforming the Approach to Racial and Religious Hate Speech Under Article 10 of the European Convention on Human Rights' (2019) 19(1) *Human Rights Law Review* 119, in particular 138–140.

speech. The reasonable speaker test is *context-sensitive*. Therefore, I will not argue for a general definition of 'hate speech'. This essay, instead, defends a *situational* approach to freedom of expression; the communicative intentions of the speaker are rationally reconstructed *case by case*, based on a reasonable interpretation of the context of utterance through a set of maxims that define communication. One point should be made clear from the outset: the reasonable speaker test is fundamental to determining when certain expressions count as hate speech, not whether hate speech is harmful or should be restricted. The test, therefore, focuses on the *linguistic and interpretive inference* in cases of hate speech, not on the *causal inference*. In other words, the reasonable speaker is understood here as an 'identifier' that *inter alia* sets standards for the plausible deniability of hate speech. It should also be emphasised that the particular version of the reasonable speaker test proposed in this essay is unprecedented and radically different from the standard American version, which is not based on Grice's theory.

My normative account of hate speech has one major advantage: it provides linguistic tools for dealing with the various layers of meaning that *prima facie* might instantiate grave forms of hate speech. I will create a taxonomy of the different forms of hate speech, based on their degree of explicitness and detachment from the literal meaning. With respect to the ECtHR jurisprudence, the most explicit forms of hate speech (e.g., the derogatory use of racial slurs, fighting words,[14] or strongly sexist remarks) will be distinguished from implicit forms of hate speech (e.g., innuendo, insinuation, and irony).

The article unfolds as follows. Section 2 shows how the lack of a precise definition of 'hate speech' paves the way for a test-based approach. To get to the core of my theoretical proposal, section 3 introduces the main conceptual tools for the pragmatic analysis of human communication, which breaks down the full meaning of an utterance into three levels: *what is said*, *what is meant*, and *what is presupposed* in an utterance.[15] Section 4 explains how the Gricean model works for hate speech. When performing hate speech, the speaker's intention is to diminish individuals on the basis of essentialised properties attributed to them *qua* members of a group. For instance, vehiculating hatred

14 I am using this expression in the sense endorsed by the US Supreme Court. See, for instance, *Chaplinsky v New Hampshire*, 315 US 568 at 571–572 (1942): '"fighting words" are those which by their very utterance inflict injury or tend to incite an immediate breach of the peace.' These words are not protected by First Amendment.

15 J Lee and S Pinker, 'Rationales for Indirect Speech: The Theory of the Strategic Speaker' (2010) 117 *Psychological Review* 785, 800–801.

on the basis of a negative property (e.g., stupidity) attributed to the person as a member of a group (e.g., ethnic minority). Section 5 turns to the pragmatic model to evaluate recent ECtHR cases on hate speech. Section 6, instead, discusses three limitations of the model: a) the absence of a European consensus, b) the presence of puzzled speakers, and c) the difficulty of determining harm. Finally, section 7 provides a summary of the main arguments and brief concluding remarks.

Before proceeding with the analysis undertaken in this study, a word of caution is necessary. *Prima facie,* the implementation of a test that was first-introduced by the United States Supreme Court – which, according to Article III US Constitution, has original and appellate jurisdiction and decides discretionarily upon the merits – to a regional Court of an intergovernmental institution, such as the ECtHR, might not be very apposite. The original mandate of the ECtHR – a body of the Council of Europe – and the US Supreme Court are quite different; the latter is certainly more complex and tangled today. This may well affect the design of the several tests applied by these courts (the three-part test, the rational basis proportionality test, the clear and present danger, and the so-called *Lemon* test, to name a few). Therefore, one might think that the ability of the ECtHR to assess a reasonable person's perception of an utterance or statement should be limited, compared to when the national courts of the Council of Europe member states – exercising their jurisdiction and margin of appreciation – assess a reasonable person's perception of the same utterance or statement. However, analysing such a conceivable limitation and, accordingly, combining the reasonable speaker's view with the margin of appreciation doctrine is beyond the narrower scope of the present inquiry. Moreover, the application of the reasonable speaker's test by the ECtHR is possible from a theoretical point of view, even if it could, under certain circumstances, lead to a dynamic, evolutionary expansion of the Court's mandate. The author of this essay is convinced that effective protection of freedom of expression today requires an understanding of the ECHR as a 'living instrument'. Finally, it should be emphasised that the reasonable speaker's test does not establish *moral* standards for Article 10 and Article 17 ECHR, as it rather deals with *linguistic data* in the form of inferential patterns that, in turn, are tied to contextual elements. In other words, the test promotes self-restraint on the assessment of an utterance or expression that *prima facie* seems to fall under one of the categories used by the ECtHR for dealing with hate speech.

2 Hate Speech: The Legal Framework and the ECtHR Approach

When we approach hate speech from a legal perspective, four different normative systems may come into play: a) the International Human Rights Law system, b) the provisions of the Council of Europe, c) the ECHR System, and, finally, d) state regulations.[16] Each level contains a wealth of norms and guidelines on hate speech. Some of these legal sources deal directly with hate speech. At the International Human Rights Law level, the OHCHR has endorsed the Rabat Plan of Action 2013, which provides a six-pronged test for the interpretation of Article 20(2) ICCPR.[17] The test takes into account: a) the social and political context, b) the status of the speaker, c) the intention to incite the audience against a target group, d) the content and form of the speech, e) the extent of its dissemination, and, finally, f) the likelihood of harm, including immediacy. The European Union has also adopted the EU Code of Conduct on Illegal Hate Speech 2017 which, although not binding, invites companies to promptly remove reported hate speech from the internet.[18] Other sources address more general issues covering particular forms of hate speech. For example, Directive 2000/31/EC (the E-Commerce Directive) regulates information society services on a general level, but several of its provisions justify the practice of filtering and removing certain types of hate speech by imposing limited liability on service providers hosting unlawful information. Looking at the EU legal system, one could also isolate other provisions that can be used to justify censorship against certain forms of hate speech. For instance, Article 1 of Council Framework Decision 2008/913/JHA of 28 November 2008 calls on states to sanction racist and xenophobic speech.[19] States have also adopted a number of autonomous measures to curb hate speech. The German Enforcement Act 2017, for example, imposes heavy fines on media outlets that fail to remove

16 The interactions between these systems – and the dialogue between the ECJ, the ECtHR, and Human Rights Committee – are very complex. See, for instance, A Frese and HP Olsen, 'Spelling it Out–Convergence and Divergence in the Judicial Dialogue Between CJEU and ECtHR' (2019) 88(3) *Nordic Journal of International Law* 429, 429–440 (on cross-references between the ECJ and the ECtHR).

17 UNGA, 'The Rabat Plan of Action' (11 January 2013) A/HRC/22/17/Add.4 Appendix.

18 European Parliament and Council Code of Conduct on Countering Illegal Hate Speech Online [2017] 128302/EU XXV.GP.

19 European Parliament and Council Directive 2000/31/EC on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market [2000] OJ L178/1 (E-Commerce Directive), which notes that civil liability applies to online platforms for failure to remove 'clearly unlawful contents', also when these contents are posted by third parties.

illegal content posted online, including forms of hate speech such as insult, incitement, and religious defamation.[20]

With respect to hate speech, the ECHR system and the International Human Rights system share several common principles. They are often analysed together, and the ECtHR also usually refers to international law provisions when interpreting the ECHR.[21] In both the International Human Rights system and the ECHR system, the obligation to combat hate speech is not *explicitly* established by specific *binding* provisions.[22] This general duty usually arises from a systematic, holistic reading of several sources, most notably Article 10(2) ECHR (limitation to freedom of expression) in conjunction with the ECHR and international law provisions, Article 19 of the Universal Declaration of Human Rights (UDHR) (freedom of opinion and expression), Article 14 ECHR (prohibition of discrimination), Article 17 ECHR (prohibition of abuse of conventional rights), Article 19(2) and Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR) (special duties and necessary restrictions to free speech), and Article 20(2) ICCPR (forms of prohibited speech). Analysis of these sources and pertinent case law reveals that the concept of hate speech

20 *Netzwerkdurchsetzunggesetz* (NetzDG, Notifizierungs-Nr 2017/127/D (Deutschland), Eingangsdatum: 27.3.2017).

21 *Loizidou v Turkey* [GC] 15318/89 (ECtHR, 18 December 1996) para 43 (the ECHR 'cannot be interpreted and applied in a vacuum'). On this point, see, M Forowicz, *The Reception of International Law in the European Court of Human Rights* (Oxford University Press 2010) 154; A Buyse, 'Tacit Citing – The Scarcity of Judicial Dialogue Between the Global and the Regional Human Rights Mechanisms in Freedom of Expression Cases', in *The United Nations and Freedom of Expression and Information: Critical Perspectives,* T McGonagle and Y Donders (eds), (Cambridge University Press 2013) 443, 446–449; AM Slaughter, 'A Global Community of Courts' (2003) 44 *Harvard International Law Journal* 191 (noting the analogies between the ECHR system and the public international law system, and the tendency to interpret the ECHR 'in harmony' with the ICCPR, even though there is no formal link or hierarchy between the two systems). For early examples of this jurisprudential trend, see, *Glasenapp v Germany* 9228/80 (ECtHR, 28 August 1986) para 48 and *Kosiek v Germany* 9704/82 (ECtHR, 28 August 1986) para 34.

22 European Commission, 'Code of Conduct on Countering Illegal Hate Speech' (2016): <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#theeucodeofconduct>, which is a non-binding commitment to remove 'illegal expressions', defined on the basis of the Framework Decision on Combatting Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, within 24 hours. Codes of conduct are generally not binding law: these measures correspond to political, programmatic, and awareness-raising objectives. The Recommendation No 20 (1997) of the Council of Europe on hate speech lacks binding force, too. See, Committee of Ministers, 'Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech"' (30 October 1997) Recommendation No R (97) 20.

itself is still largely undefined, with the lack of a definition paving the way for judicial balancing and a test-based approach.[23] The various attempts at definition pursued by non-binding EU legal sources – such as the definition of 'hate speech' in Recommendation No R (97) 20 of the Council of Europe's Committee of Ministers to member states on 'hate speech' – are also 'over-inclusive.'[24]

According to an established interpretation, such as that of *Maria Vassilari and Others v Greece*,[25] in the International Human Rights Law system, Article 20(2) ICCPR protects individuals from forms of hate speech that are not necessarily limited to violent speech: '[A]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination,[26] hostility or violence shall be prohibited by law.'[27] Article 19(3) ICCPR establishes that states are entitled to interfere with or restrict freedom of expression in order to protect other rights, and to prevent abuse of freedom of expression.[28] The principle of equality and the prohibition of discrimination are found in Article 2(1) and 26

23 R Kiska, 'Hate Speech: A Comparison Between the European Court of Human Rights and the United States Supreme Court Jurisprudence' (2012) 25 *Regent University Law Review* 107, 110: 'It is first worth considering, therefore, what "hate speech" actually is. The central problem is that nobody really knows what it is or how to define it. […] "Hate speech" seems to be whatever people choose it to mean. It lacks any objective criteria whatsoever.' For a critical review of the ECtHR jurisprudence, see, P Coleman, *Censored: How European "Hate Speech" Laws are Threatening Freedom of Speech* (Kairos 2012) in particular 115–141. The attempts of definition made by the Fundamental Rights Agency of the European Union are vague, too. See, for instance, European Union Agency for Fundamental Rights, 'Hate Speech, and Hate Crimes Against LGBT Persons' (March 2009): <https://fra.europa.eu/sites/default/files/fra_uploads/1226-Factsheet-homophobia-hate-speech-crime_EN.pdf>.

24 T McGonagle, 'Minorities and Online 'Hate Speech': A Parsing of Selected Complexities' (2012) 9 *European Yearbook of Minority Issues* 419, 422.

25 Human Rights Committee, 'Maria Vassilari et al v Greece' (29 April 2009) CCPR/C/95/D/1570/2007.

26 J Temperman, 'The International Covenant on Civil and Political Rights and the "Right to be Protected Against Incitement"' (2019) 7(1) *Journal of Law, Religion and State* 89, 89–95.

27 The 'hate speech clause' was strongly supported by the Soviet representatives. See, S Farrior, 'Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech' (1996) 14(1) *Berkeley Journal of International Law* 3, 15–17; MJ Bossuyt, *Guide to the "Travaux Préparatoires" of the International Covenant on Civil and Political Rights* (Brill 1987) 406–407; J Mchangama, 'The Sordid Origin of Hate-Speech Laws' (2012) *Policy Review* 45. I suppose that this historical fact can be explained by the close relationship between hate speech and inequality. As it is well known, communism is a form of radical egalitarianism.

28 Human Rights Committee, 'Irina Fedotova v Russian Federation' (31 October 2012) CCPR/C/106/D/1932/2010; Human Rights Committee, 'JRT and the WG Party v Canada' (6 April 1983) CCPR/C/18/D/104/1981; J Morsink, *The Universal Declaration of Human Rights: Origins, Drafting, and Intent* (University of Pennsylvania Press 1999) 66–70.

ICCPR and, as indicated above, in Article 14 ECHR.[29] There is a strong similarity between the International Human Rights Law system and the ECHR system. Nevertheless, the ultimate basis for the ECtHR's decisions on hate speech is clearly the ECHR (in particular Article 10, Article 8, Article 14 – in conjunction with Article 8 – and Article 17 ECHR). According to the prevailing view in both the International Human Rights system and the ECHR system,[30] freedom of expression is *fundamental* but not *absolute*, thus it can be balanced against competing rights and interests.[31] This doctrine differs quite markedly from the free speech paradigm prevalent in American legal culture, which is based on a libertarian approach that protects under the First Amendment extreme forms of hate speech that would not normally be protected in Europe.[32]

Within the ECHR system, the prohibition of abuse of rights (Article 17 ECHR) can be used as a pipeline to prohibit abuse of freedom of expression. Article 10(2) ECHR, too, contains a limitation clause that governments may invoke to prohibit certain forms of hate speech that are considered 'necessary' for the functioning of a democratic system. Article 10(2) ECHR provides for an exhaustive list of exceptions to the right to freedom of expression: i) national

29 The prohibition of any sort of discrimination is stated also by Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965, entered into force 4 January 1965) 660 UNTS 195 (ICERD).

30 Human Rights Committee, 'Robert Faurisson v France' (8 November 1996) CCPR/C/58/D/550/1993; Human Rights Committee, 'Rabbae, A.B.S. and N.A. v. The Netherlands' (4 July 2016) CCPR/C/117/D/2124/2011.

31 *Dudgeon v the United Kingdom* 7525/76 (ECtHR, 22 October 1981) paras 49–63; *Handyside v the United Kingdom* 5493/72 (ECtHR, 7 December 1976) para 49.

32 See, generally, A Lewis, *Freedom for the Thought That We Hate* (Basic Books 2007); GR Stone and LC Bollinger (eds), *The Free Speech Century* (Oxford University Press 2019); J Weinstein, *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine* (Westview Press 1999); J Waldron, 'Dignity and Defamation: The Visibility of Hate' (2010) 123 *Harvard Law Review* 1596, 1601–1605, showing that, in the past, the American constitutional system was often less protective towards free speech. In 1798, the Congress approved the Alien Sedition Acts of 1798, Ch 74, § 2, I, Stat 596 (expired 1801), which prohibited seditious libel. In the 1950s Supreme Court's decision *Beauharnais v Illinois*, 343 US 250, 253–57 (1952), Justice Frankfurter classified as a form of criminal libel a leaflet encouraging the protection of '"white neighborhoods" from the aggressions, robberies, rapes, and delicts of the negro.' As Waldron points out, *New York Times v Sullivan* 376 US 254 (1964) 'has removed the whole category of libel from the list of exceptions to First Amendment protection' (1607). Most decisions of the US Supreme Court reflect a marketplace of ideas approach on matters of hate speech. See, for example, *Texas v Johnson*, 491 US 397, 414 (1989); *Cox v Louisiana*, 379 US 536, 552 (1965); *Madsen v Women's Health Ctr, Inc,* 512 US 753, 763 (1994); *RAV v City of St Paul* 505 US 377 (1992) para 391 (the Supreme Court invalidates an ordinance prohibiting words 'producing anger or resentment on the basis of race', used to prosecute a man for burning a cross on the lawn of a black family).

security, territorial integrity, or public safety; ii) prevention of disorder or crime; iii) protection of health or public morals; iv) protection of the reputation or rights of others; v) prevention of disclosure of confidential information; vi) preservation of the authority and impartiality of the judiciary. The list does not mention 'countering hate speech.' This exception must, therefore, be derived by adding other premises to the list. For example, it could be claimed that certain forms of hate speech pose a threat to public safety, violate ethical values, or undermine the reputation of certain individuals and groups. It should be remembered that, in the landmark case of *Handyside v the United Kingdom*, the ECtHR held that Article 10 also protects speech that 'offends, shocks, disturbs', which, under one reading, could include hate speech.[33] An anti-militant reading of the ECHR – disconfirmed by the ECtHR practice – is, therefore, possible on paper. Be as it may, hate speech is an extremely broad concept derived from what Justice Brennan might have called the 'majestic generalities' of international law and the provisions of the ECHR.[34] The term 'hate speech' thus encompasses a whole range of expressions.[35] This broadness invites a context-sensitive analysis that checks how the expressions are used and their underlying communicative intentions.

When the Court is not applying the 'guillotine' of Article 17 ECHR (the 'abuse clause'),[36] the standard of review commonly applied to hate speech cases in ECtHR jurisprudence is the *three-part test*, which implements the margin of appreciation doctrine in hate speech cases.[37] On the one hand, this test does not include linguistics and ultimately pragmatic standards. On the other hand, the three-part test also suggests a case-by-case approach that avoids general definitions.[38] *First,* the Court determines whether the principle of legality is

33 See, for instance, *Handyside* (n 31) in particular paras 48–49, upholding a seizure of obscene books intentionally designed for children.

34 On the possibility of introducing another category of 'fear speech' as a supplement to hate speech, which operates under conditions danger of violence, see generally, A Buyse, 'Words of Violence: "Fear Speech," or How Violent Conflict Escalation Relates to the Freedom of Expression' (2014) 36(4) *Human Rights Quarterly* 779.

35 T McGonagle, 'The Council of Europe Against Online Hate Speech: Conundrums and Challenges' (Council of Europe, 7–8 November 2013): <https://rm.coe.int/168000c170f>.

36 See generally, PE de Morree, *Rights and Wrongs Under the ECHR: The Prohibition of Abuse of Rights in Article 17 of the European Convention on Human Rights* (Interstitia 2016).

37 D McGoldrick and T O'Donnell, 'Hate Speech Laws: Consistency with National and International Human Rights Law' (1998) 18(4) *Legal Studies* 453, 454. On the margin of appreciation, see, CL Rozakis, 'Through the Looking Glass. An Insider's View of the Margin of Appreciation', in *La conscience des droits. Mélanges en l'honneur de Jean-Paul Costa,* J Barthélemy and P Titiun (eds), (Dalloz 2011) 527.

38 *Müller v Switzerland* 10737/84 (ECtHR, 24 May 1988) paras 28–39.

satisfied.[39] As clarified in *The Sunday Times v the United Kingdom*,[40] restrictions on freedom of expression must be set out expressly and with sufficient precision in a written law (either a statute or a common law precedent). In the case of *Cengiz and Others v Turkey*,[41] for instance, the Turkish government was found to have failed the legality test, as Turkish law did not mention a blanket ban on an entire website as a sanction for a single piece of illegal content. The restriction was, therefore, unlawful. *Second,* the Court considers whether there are valid reasons for the restriction. This means that the restrictive measure must, on the one hand, pursue a 'pressing social need' and, on the other hand, respect the rights and reputations of others.[42] The second step is related to the legitimate aim requirement: to be valid, the restriction must pursue a legitimate aim. There is a close connection between the valid ground requirement and the list of restrictions specified in Article 10(2). Thus, in *Wingrove v the United Kingdom,*[43] the ECtHR held that the refusal to distribute a film depicting the sexual fantasies of Teresa of Avila and having as its subject the crucifixion as necessary for the protection of public morals.[44] *Third,* the Court finally considers whether the measure is necessary. The necessity threshold is not met if less restrictive and equally effective measures are available. The principle of proportionality is embedded in the third step, whereunder judges must determine whether the benefits of the ban outweigh the harm to freedom of expression caused by the restriction. *Soering v the United Kingdom* clarified that the burden of proof is on the state.[45] The three-part test has a major gap: it contains no standards for analysing the linguistic meaning of the utterance under scrutiny and no mention of the burdens for plausible deniability. This gap becomes apparent when the Court considers hate speech cases.

If we look closely at the ECtHR's jurisprudence, we can see that 'hate is a multi-faced concept used to review *content-based restrictions* on freedom of expression'.[46] The ECtHR rejects the free-market approach.[47] As noted above,

39 *Huvig v France* 11105/84 (ECtHR, 24 April 1990) paras 27–35; *Hentrich v France* 13616/88 (ECtHR, 3 July 1997) paras 10–16.

40 *Sunday Times v the United Kingdom* 6538/74 (ECtHR, 26 April 1979) paras 50–56.

41 *Cengiz and Others v Turkey* 48226/10 and 14027/11 (ECtHR, 1 December 2015).

42 *Thorgeir Thorgeirson v Iceland* 13778/88 (ECtHR, 25 June 1992) paras 60–70.

43 *Wingrove v the United Kingdom* 17419/90 (ECtHR, 25 November 1996).

44 Ibid paras 52–64.

45 *Soering v the United Kingdom* 14038/88 (ECtHR, 7 July 1989) para 111.

46 See generally, GR Wright, 'Content-Based and Content-Neutral of Speech: The Limitations of a Common Distinction' (2006) 60 *University of Miami Law Review* 333; GR Stone, 'Content-Neutral Restrictions' (1987) 54(1) *University of Chicago Law Review* 46.

47 Waldron (n 32) 1639.

the ECtHR has taken a particularist approach to assessing hate speech cases. In the case of *Lilliendahl*, the Court distinguished the 'gravest forms of hate speech' that are 'excluded entirely from the protection of Article 10' from 'less grave forms of 'hate speech' that do not fall 'entirely outside the protection of Article 10, but which it has considered permissible for the Contracting States to restrict.'[48] Among the most serious forms of hate speech is, in general, what we might call '*genocidal language games*':[49] communal practices of using dehumanising slurs and expressions of hatred for the transmission ideologies that justify genocide. Holocaust denial is a case in point.[50] The prohibition of genocidal language seems essential for the promotion of democratic values in Europe, thus substantive restrictions on the content of this kind are usually justified by Article 17 ECHR.[51] In general, the ECtHR assumes a strong correlation between hate speech and hate crimes.[52] The elimination of all forms of genocidal linguistic game is also essential for coming to terms with Europe's tragic past. The ECtHR relied on *ad hoc* assessments and held that explicit denial of the Armenian Genocide falls under Article 10(2).[53] The gravest forms of hate speech may include *justification of terrorism* and '*glorification of violence.*'[54] Justification of terrorism is often considered one of the gravest hate speech, and, even when it is not, the interest in national security generally trumps the

48 *Carl Jóhann Lilliendahl v Iceland* 29297/18 (ECtHR, dec, 12 May 2020) paras 34–35.

49 L Tirrell, 'Genocidal Language Games', in *Speech and Harm: Controversies Over Free Speech*, I Maitra and MK McGowan (Oxford University Press 2012) 174, 174–220.

50 *X v Federal Republic of Germany* 9235/81 (ECmHR, dec, 16 July 1982); *T v Belgium* 9777/82 (ECmHR, dec, 14 July 1983). The Spanish Constitutional Court Decision 235/2007 (7 November 2007) has declared unconstitutional a provision of the Penal Code that sanctioned the denial and justification of genocide because such conduct does not necessarily constitute a crime of hate (as opposed to incitement).

51 *H, W, P and K v Austria* 12774/87 (ECmHR, dec, 12 October 1989).

52 A Tsesis, *Destructive Messages* (NYU Press 2002) Chapter I; K Guenther, 'The Denial of the Holocaust: Employing Criminal Law to Combat Antisemitism in Germany' (2000) 15 *Tel Aviv University Studies in Law* 51.

53 *Perinçek v Switzerland* [GC] 27510/08 (ECtHR, 15 October 2015) paras 228–282. In this case, the Honorable Judges analytically isolate eight factors that shall be considered for those balancing cases involving hate speech: i) the nature of the statement; ii) the geographical, historical, and temporal context; iii) the extent to which the statement affects competing rights; iv) the existence of a consensus; v) the possibility that the interference can be regarded as required by the international obligation assumed by the state; vi) the method employed by the state to justify the conviction; vii) the severity of the interference; and viii) the final weighing of freedom of expression against the right to private life.

54 *Sürek and Özdemir v Turkey* (*No 1*) [GC] 26682/95 (ECtHR, 8 July 1999) paras 62–64.

freedom to explicitly justify terrorism.[55] The Court has also applied this reasoning to cases involving satire, which is sometimes considered an abuse of the right to freedom of expression under Article 17 ECHR. In *Leroy v France*, the Court thus held that a cartoon 'justifying' terrorism was not protected by Article 10 ECHR.[56]

Hate speech is often directed *against underprivileged ethnic groups and minorities.*[57] However, as the ECtHR has made clear on many occasions, 'freedom of expression constitutes one of the essential foundations of a democratic society and one of the basic conditions for its progress and each individual's self-fulfilment.'[58] Therefore, any restriction on hate speech must be understood as protecting fundamental public goods enumerated by the ECHR.[59] The ECtHR has often taken a protective stance towards unrepresented minorities. In *Zana v Turkey,* for example, the Court convicted a politician – the former mayor of Diyarbakir – for claiming that the killing of Kurdish civilians was a mere accident.[60] Even the less grave forms of hate speech that do not entirely fall outside the protection of Article 10 ECHR can be restricted by states based on Article 10(2) ECHR, which often includes cases involving *LGBTQ+ and gender discrimination.*[61] In *Vejdeland v Sweden*,[62] the Court upheld the conviction of three Swedish citizens for distributing leaflets condemning homosexual behaviour. The leaflets described homosexuality as a form of 'deviant sexual proclivity' with 'morally destructive effects.' The Supreme Court of Sweden had considered these expressions as agitation against a group, which is prohibited by Article 8 Chapter 16 of the Swedish Penal Code. In principle, Article

55 SM Boyne, 'Free Speech, Terrorism, and European Security: Defining and Defending the Political Community' (2010) 30(2) *Pace Law Review* 417, 424–430.

56 *Leroy v France* 36109/03 (ECtHR, 2 October 2008). For an overview, see, N Cox, 'The Freedom to Publish "Irreligious" Cartoons' (2016) 16(2) *Human Rights Law Review* 195, 195–221.

57 See, for instance, *Pavel Ivanov v Russia* 35222/04 (ECtHR, dec, 20 February 2007) paras 1–3.

58 *Jerusalem v Austria* 26958/95 (ECtHR, 27 February 2001) paras 69–81; *Marônek v Slovakia* 32686/96 (ECtHR, 19 April 2001) paras 337–349; *Thoma v Luxembourg* 38432/97 (ECtHR, 29 March 2001) paras 67–84; *Lingens v Austria* 9815/82 (ECtHR, 8 July 1986) para 26. For an overview, see generally, M Macovei, *Freedom of Expression: A Guide to the Implementation of Article 10 of the European Convention on Human Rights* (Council of Europe, 2004): <https://www.refworld.org/docid/49f17f3a2.html>. On 'Islamophobia', see, U Kohl, 'Islamophobia, "Gross Offensiveness" and the Internet' (2017) 27(1) *Information & Communications Technology Law* 111, 111–118.

59 Waldron (n 32) 1600.

60 *Zana v Turkey* [GC] 18954/91 (ECtHR, 25 November 1997) paras 45–62.

61 *Lilliendahl v Iceland* 29297/18 (ECtHR, dec, 12 May 2020) paras 37–45.

62 *Vejdeland v Sweden* 1813/07 (ECtHR, 9 February 2012); *H, W, P and K* (n 51) in particular para 216.

10(2) also protects statements that 'offend, shock, disturb…', but it is possible to introduce several exceptions, provided that these exceptions are construed narrowly and justified with sound arguments.[63] To come full circle, the concept of 'hate speech' is complex and tangled, thus it requires sound linguistic analysis to determine when a particular communication counts as hate speech. I argue that an updated pragmatic version of the reasonable speaker test can be useful to perform this task.[64]

3 What is Said, What is Meant, and What is Presupposed

My normative proposal is essentially based on a *Gricean account of communication*. Therefore, in this section, I will illustrate the main features of Paul Grice's model. Human communication is full of implicit, underlying information.[65] *What is said* (*explicitly*) is only the tip of the iceberg for the full meaning of an utterance.[66] Speakers take for granted a whole web of interlocking beliefs and background knowledge that help to provide full meaning of the speech act. The full meaning of an utterance is composed of at least three different levels:[67] a) *what is said*, which derives from the literal meaning conventionally expressed by an utterance; b) *what is meant*, namely, the speaker's meaning, which is a usage expressed on a particular occasion; c) *what is presupposed*, which is the background information taken for granted.

There is a fundamental difference between *what is said* and *what is meant*, which concerns the computational mechanisms for grasping each layer of meaning. The identification of *what is meant* starts from the literal meaning, but it is ultimately based on a system of Gricean maxims,[68] contextual

63 *Handyside* (n 31) para 49; *Marônek* (n 58) paras 52–62; *Thoma* (n 58) para 84; *Jerusalem* (n 58) paras 26–47; *De Haes and Gijsels v Belgium* 19983/92 (ECtHR, 24 February 1997) paras 32–49; *Jersild v Denmark* [GC] 15890/89 (ECtHR, 23 September 1994) paras 25–37; *Lingens* (n 58) paras 34–47; *Sunday Times* (n 40) paras 50–51.

64 E Camp, 'Insinuation, Common Ground, and the Conversational Record', in *New Work on Speech Acts,* D Fogal, DW Harris, and E Moss (eds), (Oxford University Press 2018) 40, 52.

65 HP Grice, *Studies in the Ways of Words* (Harvard University Press 1989) 26. See also, F Poggi, *Il modello conversazionale* (ETS Press 2020).

66 J King, 'Speaker Intentions in Context' (2014) 48(2) *Noûs* 219, 219–230.

67 HP Grice, 'Meaning' (1957) 66(3) *Philosophical Review* 377; HP Grice, 'Utterer's Meaning and Intention' (1969) 78(2) *Philosophical Review* 147.

68 HP Grice, 'Logic and Conversation', in *Syntax and Semantics 3: Speech Acts,* P Cole and JL Morgan (eds), (Academic Press 1975) 187. The idea that the identification of the full meaning of an utterance begins with the literal meaning is a fundamental assumption of Paul Grice's account. In philosophy of language, there are also more radical forms of

information, and other rules that define communication.[69] Grice's Maxims are principles that the hearers use to identify *what is meant*, and that speakers use to introduce implicit meanings. The Gricean model is based on four maxims: *quantity, quality, relation, and manner*. The *maxim of quantity* dictates being as informative as possible and giving as much information as necessary. For example, if I get lost in Berlin, and I ask someone, 'do you happen to know where Alexander Platz is?' and they simply reply, 'yes, I do', that answer would be considered a violation of the maxim of quantity because it is not informative enough for the question. The maxim of quality requires being truthful and not giving false information or information not supported by evidence. The maxim of quality is twofold: *first*, the speaker should not say what they believe to be false; *second*, the speaker should not assert anything that is not supported by adequate evidence. The maxim of relation dictates being pertinent and saying things that are relevant to the discussion. For example, if I say 'could you pass the salt, please?', and my interlocutor replies, 'I was so tired yesterday', their response violates the maxim of relation. The maxim of manner requires that one express oneself as clearly, as briefly, and as neatly as possible, thereby avoiding ambiguity. The maxim of manner is composed of four requirements: i) avoid vagueness and ambiguity; ii) avoid ambiguity; iii) be as brief as necessary; iv) be orderly and follow a systematic sequence.

A speaker may deliberately and unabashedly choose to violate (technically: *flouting*) some maxims of communication in order to convey *implicit meanings*. In the words of Nobël Laureate Daniel Kahneman:

> [c]onversations in general [...] are governed by subtle rules that determine *what is said* and *what is presupposed or implicated*. [...] When the question is ambiguous, the respondent faces the bewildering task of choosing the state of ignorance for the questioner.[70]

It is now common to use the noun 'implicature' as a term of art to denote the act of hinting, alluding, indicating, and implying a thought.[71] There are two

---

contextualism that deny the primacy of literal meaning. As indicated above, this paper advocates a Gricean model of communication. For a more radical form of contextualism, see generally, C Travis, *Occasion-Sensitivity: Selected Essays* (Oxford University Press 2008).

69 See, P Brown and S Levinson, *Politeness. Some Universal in Language Usage* (Cambridge University Press 1987) 211–227.

70 D Kahneman and MT Dale, 'Norm Theory: Comparing Reality to Its Alternatives' (1986) 93(2) *Psychological Review* 136. See also, Ibid 136; P Grice, *Studies in the Way of Words* (Harvard University Press 1991).

71 Grice (n 68) 86.

main types of implicatures: a) *conversational implicatures* (what is not conventionally implicated), and b) *conventional implicatures* (what is conventionally implicated).[72] Conversational implicatures are non-conventional, because they are essentially tied to certain situational features of the communicative exchange.[73] The speaker creates implicit meaning by flouting one of the Gricean Maxims, in a way that the listener or audience can – more or less easily – recognise. In this sense, we say that conversational implicatures must be *calculated*.[74]

Calculability (i.e., the possibility of being calculated) means that a reasonable speaker can identify implicatures through a multi-step thought process that draws on both linguistic data and situational information. Consider the following communicative exchange. (1) John: 'Would you like a Rolex?'; (2) Robert: 'I have always hated expensive watches!'. If John knows that a Rolex is expensive, then he can get the implicature: (I1) 'Robert would never want to have a Rolex', and, perhaps, also the less strong implicature: (I2) 'Robert would never want a Patek Philippe, a Breguet, a Vacheron Constantin' (some of the most expensive watches on the market). How could John get the implicature? He could follow a five-step chain of reasoning. *First,* John assumes that Robert is observing the Conversational Maxims, or at least the Cooperative Principle. *Second,* John determines that Robert has apparently violated the maxim 'be relevant'. *Third,* John can only consider irrelevance to be obvious if he assumes that a Rolex is an expensive watch. *Fourth*, John knows that Robert knows that John knows that a Rolex is an expensive watch and that he can therefore understand the implied meaning. *Fifth,* John concludes that Robert's statement implies that Robert would never want a Rolex and might never buy a Patek Phillippe, a Breguet, or a Vacheron Constantin, even though this was not explicitly stated. Certainly, the cognitive process of grasping the *implicatum* might be quick and intuitive. However, the model still provides a general idea of how we can isolate the data necessary to work out the implicature.[75] These data include: a) the conventional meaning of the words used and the identity relations between the meanings; b) the cooperative principles and the four maxims (in the example, the maxim of relation is in play); c) the context of the utterance, both linguistic and extralinguistic (perhaps John knows that Robert does not value status symbols); d) other elements belonging to the background knowledge (information about the watch market and knowledge about the most expensive brands); e) the assumed fact that the presuppositions (a-d) are

72 Ibid 41.
73 Ibid 24.
74 Ibid 30.
75 Ibid 31.

available to the communicating partners (if not, Robert would not be justified in violating the maxim).

A key property of conversational implicatures is *cancellability*: implicatures can be negated, withdrawn, disavowed, or retracted by the speaker.[76] To return to our example, imagine that John – who has understood the implicature I2 and I1 – continues by saying: (3) John: 'What a shame! I have this old Rolex that I was going to give you!'. At this point, Robert could cancel the implicature with the following utterance (4) Robert: 'I said I do not like expensive watches, but that does not mean I do not want your old Rolex as a gift!'. When we use irony, every time we use a metaphor or hyperbole, we flout the maxim of quality and produce an implicature.[77] Very often, conversational implicatures protect the speaker, who could *avoid responsibility* for their communicative intentions by simply cancelling the implicature.[78] As suggested above, there are also *conventional implicatures*, that arise from the conventional meaning of an expression. For example, if I say: 'Marco is Italian *but* he does not like football', I imply that Italian men typically do like football, and this is only suggested by the conventional meaning of the adversative 'but.' Conventional implicatures derive from semantic or syntactic features of the sentence: connectives such as 'because', 'but', and 'therefore.'

The levels of *what is said* and *what is meant* are not sufficient to map the full meaning of an utterance. A third level of meaning is needed: *pragmatic presuppositions* or, simply, *what is presupposed.* When we are dealing with implicit information, we need to separate the content that is taken for granted from what the speaker suggests, hints at, or alludes. We call the former presuppositions and the latter implicatures. Strictly speaking, presuppositions are not hints, suggestions, or clues that are inseparable from the communicative intentions of the speaker, but rather informational contents that belong to the common ground of communication.[79] The common ground of communication is a network of beliefs that are taken for granted within a communication, and that are usually shared by the participant in a communicative exchange. For example, the statement 'Riccardo quit smoking' presupposes that Riccardo was a smoker. The idea that Riccardo was a smoker is triggered by the word

76 Camp (n 64) 42; E Camp, 'Showing, Telling, and Seeing: Metaphor and "Poetic' Language"', in *The Baltic International Yearbook of Cognition, Logic, and Communication. Vol. 3: A Figure of Speech: Metaphor,* AAVV (ed), (New Prairie Press 2008) 1, 1–10.

77 Grice (n 68) 34.

78 N Asher and A Lascarides, 'Strategic Conversation' (2013) 6 *Semantics and Pragmatics* 1.

79 R Stalnaker, 'Assertion', in *Syntax and Semantics 9: Pragmatics,* P Cole (ed), (Academic Press 1978) 315; R Stalnaker, 'Common Ground' (2002) 25 *Linguistics and Philosophy* 701, 712.

'quit', which evokes the idea that something happened in the past.[80] Usually, we assume content that the participants in a communicative interaction already share. However, sometimes, we intentionally presuppose something that our interlocutors are not aware of (informative presuppositions), or something that is even false or unfounded (ideological presuppositions).

So far, we have distinguished three levels of meaning: *what is said* (the explicit meaning of an utterance), *what is meant* (implicatures), and *what is presupposed* (background content taken for granted in the utterance). As the detailed discussion in section 4 will make clear, hate speech can be both explicit – part of *what is said* – and implicit (i.e., either an implicature or a presupposition). However, from this distinction it does not follow that explicit hate speech is always forbidden and implicit hate speech is always permitted. In both cases, the outcome depends on the reasonable speaker test, as the question is whether a reasonable speaker would regard the utterance as an expression of hate, based on contextual and situational data, and filtered through the Gricean maxims. However, the distinction between the three levels of meaning is fundamental. As we shall see in due course, the techniques used for plausible deniability of *what is said* and *presupposed* differ from those used for *what is meant* and, in normal cases, denial of what is said requires the speaker to take on a heavier burden of justification.

4 The Varieties of Hate Speech and the Reasonable Speaker Test

The relationship between *what is said* and *what is meant* is highly problematic for any human rights system that allows for limitations on freedom of expression. According to a widely held intuition, we should hold speakers accountable for *what they say*, not for *the communicative intentions we ascribe to them by uncertain inferences*. Yet, we know that suggestions, allusions, and implicatures can be vehicles of hate. As we will see in this section, they can lead to hate speech. Is there a way out? A plausible solution might be to introduce a general scheme that distinguishes different layers of meaning, based on the degree of explicitness of the hate content and checks the presence of hate speech through the context-sensitive lens of a reasonable speaker at each level of meaning. These layers shall be assessed considering the linguistic uses,

80 L Karttunen, 'Implicative Verbs' (1971) 47 *Language* 340, 340–350; L Karttunen, 'Presuppositions of Compound Sentences' (1973) 4(2) *Linguistic Inquiry* 169; L Karttunen 'Presupposition and Linguistic Context' (1974) 1(1) *Theoretical Linguistics* 181.

examining the presence of focal points and contextual elements either explicitly or implicitly trigger 'hate speech' unless the speaker can plausibly deny it. Focal points are rooted in a kind of linguistic European consensus on conventional rules of communication.

As we have explained, *what is meant* is calculated through complex context-based reasoning based on normative principles (i.e., Gricean Maxims). Moreover, conversational implicatures can be cancelled. Speakers could justifiably claim that we attribute beliefs, values, desires, and intentions to them that they did not mean or think.[81] Irony and satire are a good example; a bad joke can easily be misunderstood and undone by the speaker. If the speaker legitimately invokes plausible deniability, then the piece of communication cannot be considered, strictly speaking, as hate speech, and *a fortiori* the speaker's freedom of expression should not be restricted.

In general, the communicative goal of hate speech is to diminish the target on the basis of negative properties that the target has *qua* member of the group (e.g., Martina might be the target of hate speech because she belongs to the group 'blonde women' and the speaker assumes that blondes have the essential property of being affected by cognitive impairment).[82] Hate speech is a form of vilification of the target on a social, relational, and ultimately psychological level.[83] Clear intent to disparage the target is generally considered sufficient for hate speech, even if the hate speech act misfires, and the speaker fails to induce the audience to engage in discriminatory or harmful

81 E Volokh, 'Freedom of Speech, Permissible Tailoring, and Transcending Strict Scrutiny' (1997) 144(6) *University of Pennsylvania Law Review* 2417, in particular 2418–2438; E Chemerinsky, *Constitutional Law: Principles and Policies* (Aspen Publishers 1997) 416 (strict scrutiny is the most intensive and demanding test for judicial review). See also, *Lingens* (n 58) para 42.

82 MK McGowan, 'Oppressive Speech' (2000) 87(3) *Australasian Journal of Philosophy* 389, 397–406.

83 R Langton, 'Speech Acts and Unspeakable Acts' (1993) 22(4) *Philosophy and Public Affairs* 293; R Langton 'Beyond Belief: Pragmatics in Hate Speech and Pornography', in *Speech and Harm: Controversies Over Free Speech,* I Maitra and MK McGowan (eds), (Oxford University Press 2012) 94; J Hornsby and R Langton, 'Free Speech and Illocution' (1998) 4(1) *Legal Theory* 21; R Langton, S Haslanger, and L Anderson, 'Language and Race', in *The Routledge Companion to Philosophy of Language,* G Russell and D Graff Fara (eds), (Routledge 2012) 753, in particular 754–765; R Kukla, 'Performative Force, Convention, and Discursive Injustice' (2014) 29(2) *Hypatia* 440; MN Lance and R Kukla, 'Leave the Gun; Take the Cannoli! The Pragmatic Topography of Second-person Calls' (2013) 123(3) *Ethics* 456.

behaviour.[84] Thus, hate speech such as racial slurs, ethnic epithets, burning crosses, insulting religions, inciting violence, and denigrating vulnerable persons are, to some extent, comparable to group libel.[85] However, as Waldron correctly notes,'[t]he phenomenology of this sort of assault is complex and tangled.'[86] For example, the utterance 'blacks should not have the right to vote!' uttered by a white supremacist is considered as a call to disenfranchise blacks, and the speaker's utterance may cause or reinforce discrimination, violence, and oppression.[87] Hate speech *qua* acts aimed at diminishment comprise: i) verbal aggression ('little dogs' that 'deserve a slap!');[88] ii) acts of humiliation ('hey fags – I'll buy you a free honeymoon trip to the crematorium'),[89] aimed directly at individuals and target groups; iii) acts of propaganda that promote or incite discrimination, hatred, and violence against a third person or group ('stand up against the Islamification of Belgium!').[90] Sometimes, hate speech can also be (iv) a *constitutive act of submission* (the injury is inflicted by the utterance itself) perpetrated by empowered authorities against groups or individuals (e.g., the directive 'blacks are no longer permitted to vote', issued by the South African Parliament at the time of Apartheid).[91] In all these cases, the determination of hate speech must *always* take into account the context of communication and the three layers that constitute the full meaning of an utterance.[92]

The first layer is (A) *explicit hate speech*, which refers to *what is said*. Explicit forms of hate speech include the manifest use of racial slurs, derogatory

84 A similar definitional approach was used by the Italian Court of Cassation (5th Criminal Section) in the case 44295/2005 (13 January 2005). The Court defined hate speech as: a) public; b) likely to provoke the same feelings or hatred in others; c) objectively intended to produce harmful effects.

85 Waldron (n 32) 1600.

86 Ibid 1613.

87 See generally, MK McGowan, *Just Words: On Speech and Hidden Harm* (Oxford University Press 2019).

88 *Kaboğlu v Turkey* 1759/08, 50766/10, and 50782/10 (ECtHR, 30 October 2018) paras 26–32.

89 *Beizaras and Levickas v Lithuania* 41288/15 (ECtHR, 14 January 2020) para 10.

90 *Féret* (n 10). On religious hate speech, see generally, R Moon, *Putting Faith in Hate: When Religion is the Source or Target of Hate Speech* (Cambridge University Press 2018); E Howard, *Freedom of Expression and Religious Hate Speech in Europe* (Routledge 2019).

91 I Maitra, 'Subordinating Speech', in *Speech and Harm: Controversies Over Free Speec*h, I Maitra and MK McGowan (eds), (Oxford University Press 2012) 94, 94–98.

92 R Langton, 'Subordination, Silence, and Pornography's Authority', in *Censorship and Silencing: Practices of Cultural Regulation,* R Post (ed), (Getty Research Institute 1998) 261; J Hornsby, 'Disempowered Speech' (1995) 23(2) *Philosophical Topics* 127, 127–132; Hornsby and Langton (n 83) 21–25.

words,[93] fighting words, pejorative expressions, direct insults, and other expressions that the speaker utters with the intention of diminishing a particular group or individual. Most of these linguistic uses are easy to recognise.[94] Think of expressions like 'these faggots fucked up my lunch' or 'scum!!! Into the gas chamber with the pair of them', which were uttered as Facebook comments on a post of a same-sex couple kissing and were the subject of the Second Section's judgment in *Beizaras and Levickas v Lithuania*.[95] Borrowing a metaphor first used by Kamala Harris, Elisabeth Camp has referred to these forms of speech as *bullhorns*, which everyone can understand are hate speech. As we will see in due course, the subtlest forms of implicit hate speech are more akin to *dog whistles*.

When used with the intent to offend, slurs are clear examples of hate speech: phrases like 'Mary is a dyke', 'Alex is a wop', and 'Isaiah is a kike' are semantically offensive to Lesbians, Italians, and Jews.[96] Terms such as 'dyke', 'wop', 'kike', and the like robustly ascribe negative or pejorative truth-conditional properties to individuals.[97] The pejorative perspective is absent in their neutral counterparts. The use of slurs is very often a semantic showcase that the speaker endorses the pejorative perspective, unless the speaker is being ironic, using socially accepted forms of sarcasm, or attempting manoeuvres of linguistic reappropriation.[98] When the speaker uses a slur with the clear intent to diminish or harm the target of the hate speech, the pejorative perspective persists even when the term is embedded in a question, negation, or a complex sentence ('is Mary a dyke?'). For all of these reasons, intentional slur – that is, the use of slurs that are not accepted as irony, jest, or reappropriation – is

93 See generally, J Hornsby, 'Meaning and Uselessness: How to Think about Derogatory Words' (2001) 25 *Midwest Studies in Philosophy* XXV 128.

94 See generally, L Ashwell, 'Gendered Slurs' (2016) 42(2) *Journal of Social Theory and Practice* 228.

95 *Beizaras and Levickas* (n 89) paras 10–16. The Court found that the comments violated the dignity of the applicants, in breach of Articles 8, 13, and 14 ECHR. The comments were clearly offensive and violated the psychological well-being and dignity of the couple (para 117) *qua* members of the homosexual community (para 129). In this case, the ECtHR explicitly uses the term 'hate speech' (paras 79 and 125).

96 On anti-Semitic language, see, *Pavel Ivanov* (n 57).

97 E Swanson, 'Slurs and Ideology' (Unpublished Mns 2015); A Timmer, 'Toward an Anti-Stereotyping Approach for the European Court of Human Rights' (2011) 11(4) *Human Rights Law Review* 707.

98 C Potts, *The Logic of Conventional Implicature* (MIT Press 2005) 153–193; C Hom, 'The Semantics of Racial Epithets' (2008) 105(8) *Journal of Philosophy* 416; C Hom, 'A Puzzle About Pejoratives' (2012) 159(3) *Philosophical Studies* 383; E Camp, 'Slurring Perspectives' (2013) 54(3) *Analytic Philosophy* 330.

a type of hate speech characterised by offensive autonomy and persistence. This diminishing use of slurs suggests *a strong presumption* that the speaker has uttered hate speech. Accordingly, in many contexts, the hate element underpinning intentional slurring resists plausible deniability. 'Mary is a slut, but I don't mean to insult her', uttered by a misogynistic boss to his secretary in the office, is an infelicitous speech act. It is, therefore, difficult to see how the speaker in such a case could rely on plausible deniability. Even if the boss claimed that he was only joking, any reasonable speaker would find his utterance highly offensive given the linguistic conventions that govern communicative exchanges and legitimate expectations about the speaker's beliefs.[99]

This assumption is also supported by the consideration that a *high degree of consensus* has been reached in Europe on the prohibition of these expressions. The Danish Penal Code prohibits statements 'by which a group of people is threatened, insulted or degraded on account of their race, colour, national or ethnic origin.'[100] Section 130 of the German Criminal Code, too, prohibits 'attacks on human dignity by insulting, maliciously maligning, or *defaming* part of the population.'[101] Other norms protect the dignity of individuals action attributed to the target. The state is, therefore, generally allowed to prohibit the most explicit and uncontroversial forms of hate speech. In this case the presumptions for the validity of the measures are high and the burden of justification is low. Based on these premises, a decision such as *Jersild v Denmark*, which upheld the conviction of a Danish journalist for disseminating explicitly racist remarks in his television show, with the intent to diminish and harm, is based on sound linguistic analysis.[102] Similarly, the judgment in *Beizaras and Levickas v Lithuania* is correct, from a purely linguistic perspective, in holding that terms such as 'faggot' and 'pervert' directed at gay people with the obvious intent to publicly degrade them on Facebook are clear cases of hate speech.[103]

The second layers concerns (B) *implicatures.* At this level, hate speech is expressed through veiled threats, covert insults, and other forms of *implicated*, allusive, and potentially accountability avoiding speech. If an explicit slur is a bullhorn, an insulting implicature can be likened to a dog whistle.[104] The presumption is in favour of the speaker. However, expressions that are apparently

99 A Luvell and E Lepore, 'What Did You Call Me? Slurs as Prohibited Words' (2013) 54(3) *Analytic Philosophy* 350, 353–357.
100 Section 266b(I) of the Danish Penal Code (quoted by Waldron (n 32) 1597).
101 Strafgesetzbuch [StGB][Criminal Code] Section 1.
102 *Jersild* (n 63) paras 30–37.
103 *Beizaras and Levickas* (n 89).
104 Camp (n 64) 43.

innocuous, within a specific context, might become derogatory, true threats, or incitements.[105] We have already illustrated how implicatures, when used strategically, serve to avoid an unarticulated perspective that allows plausible deniability when the speaker is prepared to deny the implicit content when it is challenged.[106] Consider the following utterance: 'You know Obama's middle name is Hussein.' Is that hate speech? Of course, 'the speaker conjures up a host of associated but unarticulated images and ideas in a way that shifts the responsibility for recovering them to the hearer, or perhaps to the broader culture.'[107] Still, one can never be sure that *what is meant* by the speaker corresponds to the racist content. As we have seen, the implicit concept (Obama is a radical Islamist) must be calculated through the Gricean Maxims and other contextual elements. The calculation can get out of hand. Moreover, the speaker may resort to cancelability as means of plausible denial. If the hearer asks, 'are you implying that Barack Obama is a radical Islamist?', the speaker can plausibly deny it. The most common forms of hate implicatures are insinuations – such as rhetorical questions – jokes, sarcasm,[108] and veiled threats.[109]

When implicatures are not tied to focal points, we lack sufficient elements to hold the speaker accountable. Camp explains, '[i]nsinuation is not a fully uniform phenomenon. It comes in degrees of obscurity, and speakers vary in their brazenness. Nevertheless, even highly transparent insinuations still admit at least some deniability.'[110] Plausible deniability is consistent with the principle *in dubio pro reo*. If we are not sure about *what is meant*, we cannot hold the speaker accountable for content that results from our risky calculation. In all standard cases (i.e., cases that do not involve reappropriation and socially accepted jokes), hatred expressed through conventional implicatures usually meets the threshold of the reasonable speaker. If a white teacher says in class 'John is black, *so* he will probably fail the exam!', this will undoubtedly be considered a racist remark by any reasonable speaker. The racist element is evoked

105 D Crump, 'Camouflaged Incitement: Freedom of Speech, Communicative Torts and the Borderland of the Brandenburg Test' (1994) 29(1) *Georgia Law Review* 1.

106 S Pinker, M Nowak, and J Lee, 'The Logic of Indirect Speech' (2008) 105(3) *Proceedings of the National Academy of Sciences* 833.

107 Camp (n 64) 42 and 46. See also, Camp 2008 (n 76).

108 E Camp, 'Sarcasm, Pretense, and the Semantics/Pragmatics Distinction' (2011) 46(4) *Noûs* 587.

109 Camp (n 64) 42 defines 'insinuation' as 'the communication of beliefs, requests, and other attitudes 'off-records', so that the speaker's main communicative point remains unstated.' See also, E Fricker, 'Stating and Insinuating' (2012) 86(1) *Proceedings of the Aristotelian Society* 61.

110 Camp (n 64) 48.

by a conventional implicature triggered by the conjunction 'so', which establishes a causal link between 'being black' and failing the exam. As explained in section 3, conventional implicatures generally cannot be cancelled. To be sure, even in this case, counterexamples can be found in which the utterance 'John is black, *so* he will probably fail the exam!' is uttered without the intention of hurting or belittling John and, therefore, the hate element can plausibly be denied. For instance, the teacher might conduct an experiment, to test students' reaction to a racist remark and – immediately after uttering it – add 'I did not mean to offend John: I just wanted to see your reactions to this highly inappropriate racist remark.' However, cases like this would be marginal in the ECtHR jurisprudence, and in most cases there is enough contextual and situational data to determine whether or not the racist implicature was genuine (e.g., the fact that the teacher had not previously made it clear that they were running a test combined with the knowledge that they are a well-known xenophobe, and the maxims would be sufficient for reading the hate *implicatum* and hold them accountable for lack of plausible deniability). Moreover, while implicatures can be *cancelled*, explicit content can only be subject to *denial*[111] (e.g., denying that the speaker uttered those words, giving a different meaning, or invoking a non-hate implicature).[112] However, the retractability of the meaning of an utterance (either by denial or cancellability) is always a matter of degree: the answer lies in contextual analysis.

Conversational implicatures are more uncertain. The performance of a punk band playing 'Punk Prayer – Virgin Mary, Drive Putin Away' in a Moscow cathedral could indeed be considered a purely ironic performance, and the band could readily cancel the element of hate.[113] In another communicative context, however, certain conversational implicatures cannot plausibly be denied. For example, a football player chanting the official salute of the Ustasha movement and the totalitarian regime of the Independent State of Croatia during a football match in front of a predominantly right-wing audience would also be understood by a reasonable speaker as ethnic hate speech.[114] Even a brief review of the facts of the case reveals the communicative intent of the football player; his political preferences were common knowledge between speaker

111 P Morency, L de Saussure, and S Osvald, 'Explicitness, Implicitness and Commitment Attribution: A Cognitive Pragmatic Approach' (2008) 22 *Belgian Journal of Linguistics* 199, 200–201.

112 R Boogaart, H Jansen, and M van Leeuwen, '"Those are Your Words, Not Mine!" Defence Strategies for Denying Speaker Commitment' (2021) 35 *Argumentation* 209, 212–216.

113 *Mariya Alekhina and Others v Russia* 38004/12 (ECtHR, 17 July 2019) paras 6–12.

114 *Simunic v Croatia* 20373/17 (ECtHR, dec, 22 January 2019) paras 45–48.

and audience, the choice of target audience was clear (right wing die-hard fans), and the player clearly knew that the audience could recognise the ideological thoughts associated with the official salute (i.e., the mutual knowledge requirement was met).

Similarly, a leaflet claiming that the Netherlands should be inhabited only by 'White Dutch People', intentionally distributed for political propaganda purposes – rather than as a grotesque fiction during an anti-fascist performance in a theatre – would leave little room for plausible deniability.[115] The implicit racist content is readily apparent given the general framework of the communicative exchange. Based on this Gricean, context-sensitive assessment, the *Norwood* case would also constitute a correct linguistic calculation of the implicatures: 'Islam out of Britain – Protect the British People', written by a far-right party on a poster leaves no room for plausible deniability, especially given the identity of the speaker and the beliefs that a reasonable person would legitimately ascribe to him.[116] The case of *Lehideux and Isorni v France*, too, was correctly argued from a pragmatic point of view. A campaign to rehabilitate the memory of Philippe Pétain cannot be considered hate speech, because the weak implicature to extreme right wing ideas could be retracted by the applicants.[117] Certainly, an outspoken defence of Philippe Pétain's political view strongly implies that the speakers are committed to far-right ideology; however, the weaker implicature that the speakers are also committed to an apology of fascism and racial crimes is based on a very uncertain inference of the interpreter. In all of these cases, analysis of situational context, combined with background information about the speaker's beliefs and Gricean maxims determines the outcome.

The third category includes (C) *hate speech by presuppositions*. Presuppositions are tacit assumptions, beliefs, and attitudes that are taken for granted. Hearers either take these background elements for granted because they already know and share the speaker's presuppositions, or they accommodate the common ground of shared beliefs. The utterance 'No to the gyspsification of Bulgaria!'[118] is an example of hate presupposition. The term 'gyspsification' typically presupposes a negative attitude towards the Roma people, and this was clearly mutual knowledge between speaker and hearer because the pejorative use is, in turn, tied to a focal point. Through this presupposition, the speaker

115 *Glimmerveen and Hagenbeek v the Netherlands* 8348/78 and 8406/78 (ECmHR, dec, 11 October 1979).

116 *Norwood v the United Kingdom* 23131/03 (ECtHR, dec, 16 November 2004).

117 *Lehideux and Isorni v France* [GC] 24662/94 (ECtHR, 23 September 1998).

118 *Behar and Gutman v Bulgaria* 29335/13 (ECtHR, 16 February 2021) para 14.

introduces a contemptuous attitude toward Roma into the common ground of communication.[119] Malicious speakers might strategically exploit these presuppositions to circumvent censorship. Consider the following *standard* example of an offensive presupposition. Two Italians are accused of raping a woman in a London club. The defendants' lawyer states: 'These are boys from good families, who, like everyone else their age, responded to the woman's unequivocal advances.' This statement presupposes that the victim made explicit advances and *a fortiori* that the woman consented to sexual intercourse. This assumption is triggered by the phrase 'unequivocal advances', and suggests that the woman was not actually raped. Viewing victims as culprits of a rape crime is clearly a gender discriminatory notion that creates a degrading attitude towards women.

To avoid excesses of censorship and pragmatically unsound legal arguments that arbitrarily punish speakers, judges could return to the *reasonable speaker test* and consider every level of the utterance. In US constitutional law, this test is generally applied to cases of workplace harassment, libel, incitement, extorsion, true threats,[120] and bribery cases.[121] The reasonable speaker test specifies that something *counts as* sexual harassment, incitement, extorsion, or bribery *only if* a reasonable speaker would understand the utterance to be harassment, libel, incitement blackmail, true threat, or bribery.[122] For example, if a reasonable speaker would take a *prima facie* threat to be a clearly hyperbolic speech, then the speech is protected.[123] What is more, we can recast the reasonable speaker test using the Gricean account for calculating the full-meaning of the utterance through its three layers. The innovative Gricean version of the reasonable speaker test proposed in this essay yields significant payoffs. Unlike the standard US version of the test,[124] the pragmatic approach ties the speaker's communicative intention to objective linguistic conventions that govern normal communicative exchanges and to inference patterns that process contextual data.[125] According to Grice's seminal analysis, the speaker is accountable when there is no room for plausible deniability or *quid pro quo* given the

119 On the non-derogatory uses of the 'N-word', see generally, K Randall, *Nigger: The Strange Career of a Troublesome Word* (Vintage 2002).

120 *Chaplinsky* (n 14).

121 E Smith, 'Freedom of Speech and the Classification of True Threats' (2015) 2(1) *The Cohen Journal* 1, 3–12.

122 Camp (n 81) 52.

123 *Watts v United States* 394 US 705 (1969).

124 Ibid.

125 *United States v Fulmer* 108 F.3d 1486, 1491 (1st Cir 1997).

linguistic conventions that govern communicative exchange.[126] The focus is on the speaker, who must foresee that his utterances could be understood as serious expressions of hatred that cause harm under a normal negligence standard.[127] Something is worth repeating: the function of this test is purely classificatory, and precedes further determination of harmful effects. In other words, the test provides a tool to assess whether the interpretive inference of the ECtHR is appropriate. The causal inference shall be subject to a different test.

Under the Gricean model, the full meaning of utterances is never decided solely by their literal meaning. This reasoning also applies to explicit hate speech. Even in this case, the decision-maker must examine the presence of conventional or conversational implicatures. The implicit meanings of an utterance are highly sensitive to the context. This consideration also applies to slurs. The use of a slur usually triggers hate speech; however, sometimes, slurs can go through a process of *reappropriation* and be transformed into positive or friendly qualifiers. For instance, 'you're a slut' can have very different meanings, depending on *who* says it and in *what* context. Thus, to classify a slur as hate speech, one must consider both the implicatures and the literal meaning, based on the speaker's reasonable expectations. In summary, the full meaning of hate speech always depends on the combination of what is said, what is meant, and what is implied, never on the literal meaning alone. That said, the commitment denial of explicit slurs requires a *higher burden of proof*, as the speaker cannot turn to cancellability. Instead, they shall demonstrate the presence of an indirect speech (e.g., show that an innocent joke was mistaken for an insult), or point to the exceptional circumstances that characterise the context of utterance.[128] Judicial application of the reasonable speaker test promotes a tailored approach to communication rather than blind content-based general restrictions, that could ultimately silence legitimate expression, as is the case with the purely syntactic filters endorsed by social networks and online service providers for content moderation. On several occasions, these filters have resulted in innocuous remarks being deleted. For example, social networks have silenced drag queens because

126 Camp (n 81) 52; E Camp, 'Conventions' Revenge: Davidson, Derangement, and Normativity' (2016) 59(1) *Inquiry* 113.

127 PT Crane, '"True Threats" and the Issue of Intent' (2006) 92 *Vanderbilt Law Review* 1225, 1225–1227.

128 Boogaart, Jansen, and van Leeuwen (n 112) 212–216.

they believed that permissible and innocuous forms of reappropriation were a form of 'toxic' speech.[129]

The dividing line between hate speech based on conventional implicatures and hate speech based on conversational implicatures, too, is narrow, and must be determined in light of the context, by a plausible reconstruction of the speaker's beliefs. The latter can be cancelled, but the former cannot. In real cases, however, both types of implicature are often at play simultaneously. Consider again the example of the white teacher who says 'John is black, *so* he will probably fail the exam.' In the normal case, this would undoubtedly be considered a racist remark by any reasonable speaker, because of the conventional implicature triggered by the conjunction 'so.' However, only by also considering conversational implicatures and presuppositions can one decide whether this is actually *what is meant*, or whether the teacher is in fact being sarcastic by, for example, mocking the students' supposed racial bias, or critical by, for example, denouncing what they think are racial assessment standards that they must apply. In understanding the implicit dimension, context – that is, 'the speaker and the hearer, their common perceptible environment, their previous utterances, and all of their relevant beliefs' – plays a central role.[130] Thus, the reasonable speaker test requires the decision-maker to make a *sound inference to the best interpretation*.[131] The best interpretation of an utterance is the logically strongest proposition that is consistent with what is non-controversial in a communicative exchange between hearer and speaker regarding the utterance in a given context, and is consistent with the common ground between speaker and hearer.[132] It is not always easy to determine the speaker's commitment to the communicative content. In many cases, however, there are certain focal points that make a particular use of language crystal clear (within a particular context). These focal points, which map equilibria in meaning use, are constraints on plausible deniability.[133]

129 T Dias Olivas, D Antonialli, and A Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 *Sexuality and Culture* 700.

130 AP Martinich, 'Conversational Maxims and Some Philosophical Problems' (1980) 30 *Philosophical Quarterly* 215, 221.

131 JD Atlas and SC Levinson, '*If*-Clefts, Informativeness, and Logical Form: An Introduction to Radically Radical Pragmatics (Revised Standard Version)', in *Radical Pragmatics,* P Cole (ed), (Academic Press 1981) 1.

132 JD Atlas, *Logic, Meaning and Conversation* (Oxford University Press 2005) 95.

133 Lee and Pinker (n 15).

The reasonable speaker test takes into account the three levels of meaning, context, and shapes sound expectations of mutual beliefs shared in a communicative exchange to determine what our words say vis-à-vis what the speaker implies by using them in a particular way.[134] To make this assessment, several contextual elements must be viewed through the lens of a reasonable person (i.e., the 'man on the Clapham omnibus', or the reasonable bystander),[135] who is fair minded, informed, competent in the language of the communicative exchange, and possesses ordinary language skills, necessary for working out implicatures.

The commitment of the speaker is determined by specific *inference patterns* that take into account the following: i) the explicit meaning of the words; ii) the three levels of meaning, as derived from a reasonable understanding of the utterance in light of the Gricean maxims; iii) the linguistic and situational context of the communication, focusing on the interaction 'speaker-audience-background of the communicative exchange', which includes the domain of discourse, the general topic, and the social relations between the parties; iv) the predictability and severity of the unlawful outcomes; v) the extent of the speaker's knowledge or expectation of harmful effects, and the presence of a possible motivating goal to cause harm, humiliation, violence, and so forth; vi) the presence of negative appraisals of group members and possible action tendencies of revenge, social exclusion, or personal attack; vii) the presence of afterthoughts or immediate correction of the message, and the possibility of reappropriation phenomena; viii) the possibility that the utterance is an example of artistic, literary, scientific, satirical, and ironic expression; ix) the presence of reckless behaviour on the part of the speaker; x) the condition that the speech stigmatises the target by attributing to it extremely negative essential characteristics. These elements lead both to contextual enrichment of the meaning of an utterance and to the calculation of implicature, thus, they must be taken into account by the ECtHR before upholding restrictions on hate speech.[136] Generalising, the test establishes the speaker's commitment to the hate content based on the degree of explicitness of the meaning conveyed and the availability of plausible deniability. Plausible deniability, in turn, depends on the strength of the implicatures and presuppositions: the stronger

134 P Grice, 'Reply to Richards', in *Philosophical Grounds of Rationality,* R Grandy and R Warner (eds), (Claredon 1986) 45, 59.

135 *Southern Foundries (1926) Ltd v Shirlaw* [1939] 2 KB 206 (17 March 1939).

136 See generally, R Carston, *Utterances and Thoughts: The Pragmatics of Explicit Communication* (Blackwell 2002).

the implicature/presupposition, the higher the speaker's degree of accountability and burden of deniability.

5 A Gricean Assessment of the ECtHR Recent Case Law

I will now apply the pragmatic account to a handful of recent landmark cases on hate speech, in order to determine whether they were correctly decided. Let us start with *Ottan v France*, a case concerning hate speech against ethnic minorities.[137] According to the pragmatic model developed in this essay, *Ottan* was an easy case, based on sound pragmatic analysis. The applicant is a lawyer who received a disciplinary sanction after commenting on the acquittal of a police officer who had killed a young man of foreign origin. The lawyer simply said that the acquittal was no surprise, because the jury had been 'white, exclusively white'. The ECtHR found that the disciplinary measure violated Article 10 ECHR. From a semantic point of view, the term 'white' has no negative racial connotation. It is neither a slur nor an insult. Moreover – as the Court rightly argues, taking into account both the communicative context, the apparent communicative intention of the lawyers, and the broader background of the case – the applicant's remark does not imply any form of ethnic hatred against the (alleged) target group (i.e., white people). What is more, the lawyer belonged to that group. In a more pragmatic lexicon, we would say that the attorney's remark does not trigger any connotation of hatred, and it seems quite difficult to isolate a (weaker) conventional implicature beyond the (strong) suggestion, or allusion, that the jury was biased. The lawyer has clearly flouted the Maxim of manner for vehiculating the strong implicature 'the jury was biased.' Any reasonable speaker would understand that. Given the context of the communicative exchange, the implicature 'the jury is biased' cannot be convincingly cancelled. Indeed, the applicant does not even attempt to pursue this line of defence (namely, to invoke plausible deniability for the claim of bias), nor does it seem necessary to do so. No reasonable speaker would ever consider a discussion of a jury's 'racial bias' – an incredibly fruitful topic of scientific inquiry – to be hate speech. Furthermore, given the current speech practices in Europe, being white does not presuppose a negative judgment of that ethnic group. Therefore, the phrase 'white, exclusively white' does not trigger hate speech related to the background of the communication. The lawyer's speech was clearly not aimed at mistreating or humiliating white people, thus

137 *Ottan v France* 41841/12 (ECtHR, 19 April 2018) paras 49–79.

the essentialisation element that characterises hate speech was not met. The case might have been different if a non-white applicant had asserted, 'Jurors were white, exclusively white, and you know what this race did during World War II.' In such a case, the reasonable speaker test suggests the existence of a hate implicature. However, the speaker's action tendency of social exclusion, possible motivational goal of harming or general racist behaviour towards white were absent in the situational context.

Let us now examine a case of 'sex-related hate speech', *Vejdeland v Sweden*.[138] According to the pragmatic model, this case was correctly decided, but the inference to the best interpretation was insufficiently argued. The Court's reasoning should have developed the linguistic analysis of the expressions in question. Let me elaborate. The expressions in the leaflets posted in the secondary school lockers do not qualify as *explicit* hate speech (i.e., hate speech that relates to the dimension of *what is said*). As the applicants pointed out, the language did not contain sexist slurs, overt incitement, and the like. The flyers criticised 'homosexual propaganda', and referred to homosexuality as a 'deviance' with 'morally destructive effect' that played a role in the spread of HIV/AIDS and paedophilia. Was this a form of *implicit* hate speech? The applicants maintained that they were only concerned to stimulate discussion in schools, not to denigrate. In slightly more technical jargon, we would say that they *claimed plausible deniability*. In particular, they tried to cancel the implicature and deny the intention of diminishing and inciting harm against homosexuals. The Court held that Article 10 ECHR was not breached, because the suppression of their speech met a *legitimate aim* – the protection of the 'reputation and rights of others' – and that the interference was *necessary in a democratic society*. But this consideration begs the question. *First*, the hateful element implicit in these claims must be identified by calculation, as a conversational implicature, or by the presence of a lexical element that activates a presupposition or conventional implicature. In other words, the threat and incitement of hatred against homosexuals must be demonstrated through contextual linguistic analysis, by showing that the specific use of these expressions by the applicants presupposes or implicates an element of hatred. It turns out that this evidence was available, and that the Court could have illustrated how any reasonable speaker would assume that the expressions used in the flyers presuppose or suggest a hate content. *First*, the identity of the speaker is significant. The fact that the leaflets were printed by the 'National Youth' – a far-right homophobic political group – could have played a key role in elaborating

138 *Vejdeland* (n 62).

the hate implicature (e.g., attributing discriminatory or malicious attitudes to the member of the political group). *Second,* the consideration that they were directly addressing youngsters, rather than parents and teachers, shows that the requirement of sincerity is not met, as the applicants' communicative act only *purported* to promote discussion, but the aim of their speech act was to *persuade* (i.e., to stir up prejudice on highly influential issues). *Third,* the general cultural context in Europe, where homophobic and transphobic sentiments are still widespread, is also fundamental to the activation of a hateful condition. The uptake of the 'hidden' hatred would probably have been effective, and this was predictable. *Fourth,* looking at the background of the communication, the term 'defiance' and the idea of being a cause of paedophilia and HIV triggers pejorative and derogatory thoughts *via* essentialisation of negative qualities. Any reasonable speaker would read this presupposition, and the context, coupled with knowledge of the speaker's political views, as hardly allowing for the possibility that the utterance is merely a joke, let alone a form of linguistic reappropriation. I argue that this form of pragmatic analysis could lead to a 'stronger reasoning' that – as the concurring opinions of Spielmann and Nussberger point out[139] – reveals a *hidden* aim of insult and denigration behind the *apparent* aim of furthering the discussion. *Vejdeland* still has another spot: the ECtHR made no real effort to establish the intent to harm caused by the speech (namely, the foreseeability that secondary school students would engage in transphobic or homophobic bullying after reading the leaflets) and the recklessness of the speaker. Further arguments were required in this regard and they were available. Still, the implicature suggesting hatred against homosexuals is strong and the hate *implicatum* is clearly what a reasonable speaker would recognise in this context. The ECtHR correctly follows the precedent of *Féret v Belgium*, which upheld a man's conviction for incitement to hatred and discrimination (i.e., hate speech) for distributing leaflets promoting racism and discrimination during elections. From a purely linguistic point of view, the cases are quite similar; in *Féret*, the racist and xenophobic intent was clearly in Mr Féret's ideology (he was a radical member of the Front National) and in the literal meaning of his statement (which explicitly advocated racial segregation, unequal treatment of foreigners, the eradication of Islamic culture, and the suspension of fundamental rights for non-European residents of the 'Cuscus clan').[140] However, there is a major difference between the two cases. Mr Féret did not even try to deny his xenophobic feelings and

139 Ibid Concurring Opinion of Judge Yudkivska joined by Judge Villiger.
140 *Féret* (n 10) paras 7–18.

intention to disparage the target. Apart from the determination of harm, Mr Féret's remarks were a clear case of explicit hate speech from a pragmatic perspective. In both *Vejdeland* and *Féret,* the implicatures were very strong and easily accessible.

Let us now analyse a seminal online hate speech case, namely, *Delfi v Estonia*,[141] which deals with corporate liability for user-generated comments on the internet. In balancing privacy and freedom of expression, the Grand Chamber comes quite close to the approach of a reasonable speaker, by pointing out that the comments in question ('bloody shitheads', 'a good man lives a long time, a shitty man a day or two') were generally perceived as demeaning and defamatory on a national level and were – in terms of language – clear insults and threats.[142] The element of hatred was present at the level of *what is said*, for they were expressions that 'did not include sophisticated metaphors or contain hidden meaning or subtle threats.'[143] The burden of justification in this case was low. The denial did indeed fall under the speaker's responsibility, but the inference to the best interpretation of the Court was clearly sound. The commitment of the speaker towards the explicit meaning was legitimately inferred by the Honourable Judges; no further plausible denial or retraction of the content was available. Any reasonable speaker would recognise the hate tinge and the presence of a threatening implicature, since this use of language in context were tied to focal points. Denial was thus *possible*, but not *plausible*, since the use of these insults is conventionally tied to language use within the speaker's community of reference. The Grand Chamber was right. Sophisticated linguistic analysis does not even seem to be necessary in this case, because the utterances are *manifestly* unlawful, even in light of the context,[144] which left no room for the denial of the literal meaning, or the claim that it was a slip of the tongue. Moreover, the restriction was justified by the legitimate aim of protecting the reputation of others.[145] The Court's reasoning is correct but it has a blind spot. As in *Vejdeland*, the Grand Chamber's ruling overshadows the finding of the harmful intent, which is mentioned only *en passan.*[146]

Let us now briefly consider *Ibragim Ibragimov and Others v Russia*,[147] a case involving the glorification of violence. This case was also correctly decided.

141 *Delfi AS v Estonia* 64569/09 (ECtHR, 10 October 2013).
142 Ibid para 114.
143 Ibid para 156.
144 Ibid para 117.
145 Ibid para 63.
146 Ibid para 157.
147 *Ibragim Ibragimov and Others v Russia* 1413/08 and 28621/11 (ECtHR, 28 August 2018).

As the ECtHR found, the moderate and non-violent image of Islam defended by Said Nrusi qualifies as hate speech. No explicit hate speech was used (i.e., hate speech did not originate at the level of *what is said*), and Russia had not convincingly demonstrated the presence of hate implicatures or presuppositions in Nrusi's language, as: in the context of a religious book, it is not a crime to call non-Muslims 'dissolute' or 'idle talker.' Moreover, these words do not implicate an incitement to violence against non-Muslims.[148] The inference that leads from 'Said Nrusi said non-Muslim are "dissolute"' to the implicature 'Said Nrusi obviously means that non-Muslims can legitimately be targets of violence, freely diminished in public, and so on' is not the best interpretation of the speaker's communicative intent. A reasonable speaker would not recognise the presence of a hate colouring, especially given the specific context: Said Nrusi has offered a standard interpretation of the Qur'an. Moreover, the Mr Nrusi has not openly violated any of the maxims for conveying implicit content. He did not violate manner, quality, quantity, or relation. In other words, he was sincere, he offered a sufficiently detailed moral argument for his conclusion, the moral condemnation was clear, and appropriate in light of the communicative exchange.

Up to this point, we have essentially considered cases that meet the thresholds of the reasonable speaker test. However, the ECtHR has sometimes departed from a pragmatically sound analysis. Let us compare *Ibragim* with *Nix v Germany*,[149] which was, instead, wrongly decided in light of the pragmatic model. The use of the image SS Chief Heinrich Himmler in a blog as a tool to criticise a 'slimy staff member' of the employment agency who allegedly engaged in ethnic profiling clearly presupposes a negative attitude towards National Socialism on the part of the speaker. While the ECtHR correctly found that this use is not hate speech, the Court failed to recognise that the applicant's conduct – i.e., turning to Himmler's image to denounce discrimination against his German-Nepalese daughter – was clear evidence that the applicant rejected Nazi ideology.[150] Any reasonable speaker would read that presupposition. Moreover, the decision develops no discussion of intent to harm and no reconstruction of the general communicative context in contemporary Germany, where well-known newspapers, satirical magazines, such as *Titanic,* and blogs have repeatedly used images of Hitler and swastikas without sanction. The post did not show any serious intention to hurt, humiliate, or attack the employee, and the negative evaluation of his behaviour was not tied to

148 Ibid paras 114–121.
149 *Nix v Germany* 35285/16 (ECtHR, dec, 13 March 2018).
150 Ibid paras 51–54.

any essential characteristic that the target (i.e., the staff member) exhibited *qua* member of a particular group (e.g., Germans or German civil servants). The speaker clearly resorted to hyperbole and exaggeration against discrimination. Thus, the essentialisation characteristic of hate speech was completely absent. Also, the perception of the relationship between the author of the post and the target does not indicate that the speech act was aimed at undermining the basic social position of the civil servant. Rather, it was the overt act of denouncing discriminatory behaviour. There were no other implicatures conveying thoughts of violence, revenge, humiliation, or social exclusion. Finally, the behaviour of Mr Nix does not indicate reckless or malicious intent, let alone hostility or a propensity for violence. In sum, from a pragmatic point of view, the case of *Nix* is an example of misattribution of commitment.

Marais v France, too,[151] clearly does not meet the reasonable speaker test. The three-page article denying the existence of a gas chamber at Stuthof-Natzweiler cannot be classified as Holocaust denial. *First*, the article, written by Mr Pierre Marais, does not explicitly deny the Holocaust (it is just the opposite). Moreover, it does not contain any slurs or derogatory words against the Jews. *Second,* the article does not deny the existence of the Stuthof-Natzweiler, a concentration camp, and explicitly describes the research results as 'not aspiring to scientific precision' and, most importantly, as an exception (i.e., a 'special case'), which *presupposes* that – according to the author – gas chambers were *normally* present in concentration camps (which is exactly the opposite of Holocaust denial). This presupposition undermines the implicature of the more general assertion 'gas chambers in general were technical improbable', miscalculated by the Court. The non-literal reading that yields a diminishing implicature or a denigrating effect was not backed up by linguistic focal points. In other words, from the perspective of a reasonable speaker, the applicant has plausibly denied the more general implicature by relying on the 'special case' argument. The context is clear enough. Mr Pierre Marais' use of language did not implicate forms of contempt, disgust, anger, nor a clear motive to harm and humiliate the victims of the Holocaust, nor a desire for revenge. There were no other considerations of his social status to suggest that the above remarks involved mistreatment or humiliation, or that the purpose of his speech was to undermine the social standing of certain ethnic and social groups. The Court did not isolate sufficient contextual assumptions for this inference to the best interpretation. The domain of discourse was restricted to a very specific location (not concentration camps in general), and the means chosen to convey

151 *Marais v France* 31159/96 (ECmHR, dec, 24 June 1996).

this idea (a low impact essay, rather than a more effective propaganda devise) suggests that, contrary to what the ECtHR considered, the motivating goal of undermining the fundamental social standing of Holocaust victims was not present, contrary to the ECtHR's view. Any reasonable speaker would understand these points. To come full circle, this case is also a clear example of false attribution leading to unwarranted conclusions.

6 Limitations of the Model: Absence of European Consensus, Puzzled Speakers, and Difficulty in Harm Determination

The reasonable speaker standard provides a possible perspective for the inference to the best interpretation in hate speech cases (a 'filter' for reading linguistic phenomena), not a knock-down answer to every challenge to hate speech. Notably, the reasonable speaker test leaves room for limited judicial discretion and context-sensitive judgments. Application of the reasonable speaker test requires a multifactorial, context-dependent analysis that takes into account *inter alia*: a) the conventions underlying irony, satire, and hyperbolic language; d) the use of Grice's maxims; and, c) the plausible perception of the speaker's beliefs. Thus, hard cases will often arise, especially when there are profound disagreements or essentially contested moral concepts between the speakers.[152] In particular, the Gricean version of the reasonable speaker test has three limitations: a) the presence of puzzled speakers, b) the lack of a European Consensus, and c) uncertainty about the causal inference.

Let us consider the possibility of 'puzzled speakers'. For certain expressions, there could be linguistic disagreement as to whether they qualify as 'hate' speech. This is due to the cultural, social, religious, and political diversity of the European context. However, *linguistic* disagreements should not be equated with *moral* disagreements. A reasonable speaker might understand that an utterance constitutes hate speech even when they believe that they are morally justified in using it, because they perceive the target as 'inferior.' This understanding is precisely the reason for using the hate colouring rather than a neutral counterpart. In such a case, the speaker is held accountable regardless of their moral beliefs. Even with this caveat, some cases are borderline. *Leroy v France*, for example, is a hard case. Does a vignette drawn by a Basque cartoonist, depicting the collapse of the World Trade Center, along with the slogan 'We

152 WB Gallie, 'Essentially Contested Concepts' (1955/1956) 56 *Proceedings of the Aristotelian Society New Series* 167.

all dreamed of it…Hamas did it' count as condoning terrorism and glorifying violence? It is clearly a form of implied speech, as the ECtHR acknowledges ('allusion').[153] Can the hate implicature be cancelled by the speaker? Can the cartoonist plausibly deny it? These are difficult questions.

A second limitation of the model is the lack of European consensus on certain forms of expression (B). As Panos Kapotas and Vassilis P Tzevelekos point out, "[c]onsensus' is 'a riddle, wrapped in a mystery, inside an enigma';[154] the ECtHR acts as 'a barometer of evolution […] within the European continent', mapping the existence of 'pan-European standards' shared by the High Contracting Parties.[155] States may disagree, however, and national jurisdictions may also be divided. The absence of European consensus suggests a greater margin of appreciation of national authorities.[156] Thus, if there is strong and widespread disagreement about what constitutes hate speech, the ECtHR should advocate *self-restraint* and protect the pluralism of the European society, on the one hand, and the rights of political minorities on the other.[157] France, for example, passed a law requiring media outlets and platforms to remove 'manifestly unlawful' hate speech within a few hours of notice. The content of this law was similar to the German Network Enforcement Act, which is widely accepted in Germany, but the France Constitutional Council overturned this law[158] on the grounds that it was a disproportionate and unnecessary restriction on freedom of expression. There is also a lack of sufficient consensus among European states on key issues, such as Holocaust denial. Again, there is only an incompletely theorised agreement on the need to combat negationism,[159] but no common standards for implementing this goal in individual

153 *Leroy* (n 56) para 42.

154 P Kapotas and VP Tzevelekos, 'How (Difficult Is It) to Build Consensus on (European) Consensus?', in *Building Consensus on European Consensus: Judicial Interpretation of Human Rights in Europe and Beyond,* P Kapotas and VP Tzevelekos (eds), (Cambridge University Press 2019) 1, 1. On European consensus, see also, K Dzehtsiarou, *European Consensus and the Legitimacy of the European Court of Human Rights* (Cambridge University Press 2015) 39–56, and LR Helfer, 'Consensus, Coherence and the European Convention on Human Rights' (1993) 26 *Cornell International Law Journal* 133, 133–136.

155 Kapotas and Tzevelekos (n 154) 4–5.

156 Ibid 5.

157 VP Tzevelekos and K Dzehtsiarou, 'International Custom Making and the ECtHR's European Consensus Method of Interpretation' (2016) 16 *European Yearbook on Human Rights* 313; N Krisch, 'The Open Architecture of European Human Rights Law' (2008) 71(2) *The Modern Law Review* 183.

158 Decision No 2020–801 DC of 18 June 2020.

159 CR Sunstein, 'Incompletely Theorized Agreements' (1995) 108(7) *Harvard Law Review* 1733.

cases.[160] If the ECtHR is unable to find a sufficient level of consensus through a comparative analysis of the common practices in Europe,[161] the least intrusive option is to protect freedom of expression. The lack of European consensus is also relevant to cases decided primarily under the rubric of Article 17 ECHR. The militant demand that the traditional neutral model of liberal democracy shall be abandoned in favour of systems that 'protect' democracies by restricting 'extreme expression' generally presupposes at least a partial consensus on what counts as extreme discourse. Of course, the variable 'extreme expression' is context-dependent and allows for cross-national variation. Therefore, it is of central importance to determine the degree of European consensus.

A third fundamental limitation of this model is the *uncertainty about the causal link*. We have already explained that the reasonable speaker test is a tool for identifying hate speech, not a standard for determining the presence of harm, or the subsistence of a causal link between speech act and harmful consequences. This is a limitation of our model, because detecting harm and justifying a causal inference is not smooth sailing. If effective, hate speech manipulates the coalitional psychology of victims and bystanders. Victims are belittled or demeaned, whilst bystanders are asked to support this imbalance with their consent, reinforcing the behavioural mechanisms that cause further harm (e.g., acts of violence or humiliation). However, as Eric Heinze correctly notes, harm is clearly not inherent in hate speech,[162] and hate speech is not harmful per se. The ECtHR should therefore use sound statistical reasoning to determine whether there is a causal link between harm and speech. However, the causal inference is too often the product of guesswork, subjective intuitions, spurious correlations, and subjective biases. At worst, the causal connection is merely symbolic, 'metaphorical',[163] or based on a 'purely rhetorical consequentialism.'[164] The attribution of harmful effects to the Nazi propaganda in the 1930s, for example, is based on sound reasoning. However, we must avoid what Heinze calls 'the Weimar fallacy' and generalise causal correlations that

160 See generally, D Kagiaros and VP Tzevelekos, 'The Importance of State Practice in the Shaping of International Standards Pertaining to the Clash Between Free Speech and the Banning of Negationism: The Contribution of the Greek Legal Order', in *Responsibility for the Denial of International Crimes,* P Grzebyk (ed), (Instytut Wymiaru Sprawiedliwośki 2020).

161 K Dzehtsiarou, 'Comparative Law in the Reasoning of the European Court of Human Rights' (2010) 10 *University College Dublin Law Review* 109, 139.

162 Heinze (n 7) 9.

163 Ibid 33.

164 Ibid 33 and 97.

are deeply context-specific.[165] Spurious causal correlations and incorrect generalisations affect some of the prominent deontological and dignity-informed theories of hate speech. According to Heinze – with whom I agree on this point – fallacies and biases should be replaced with empirical evidence and statistically sound reasoning.[166] Heinze's argument highlights the inevitable difficulties in determining the causal links between hate speech and harm, which are not resolved by the reasonable speaker test, which has a more limited purpose, namely to determine whether a contextualised utterance counts as hate speech, from a purely linguistic and, ultimately, pragmatic perspective. Nor are the traditional American standards for reviewing free speech cases very helpful on this point. The statement that the right to free speech can be restricted if there is a clear and present danger does not resolve the puzzle, as the term 'clear and present danger' is clearly vague.[167] Moreover, excessive restrictions could lead to more violence and polarisation, eventually resulting in an erosion of democratic institutions. If Heinze is right, as I believe, then the ECtHR, too, should support the presumption of liberty and not impose further restrictions on hate speech unless harm is demonstrated.[168] Slurs, fighting words, and insults can also have harmful effects (namely, inequalities among the targets or acts of violence perpetrated by the audience).[169] The harm caused to the target can be either direct (anxiety, fear, and moral damage) or indirect (actual discriminatory practices or political disempowerment of a particular group), and both long-term (mental illness) and short-term (an act of violence against a member of the target).[170] It may also include loss of credibility due to prejudice related to social identity.[171] However, the causal inference must be based on sound causal links and, in many cases, the ECtHR has overlooked issues of causality by upholding a general presumption of harmful effects or claiming a dubious causal link between expression and harm through the likelihood test. For example, in *Soulas and Others v France*, the Fifth Section

165 Ibid 131–132.

166 Heinze (n 7) 126–128.

167 *Schenck v United States* 249 US 47 (1919).

168 D Brink, 'Millian Principles, Freedom of Expression, and Hate Speech' (2001) 7(2) *Legal Theory* 119, 130–157.

169 I Maitra and MK McGowan, *Speech and Harm: Controversies Over Free Speec*h (Oxford University Press 2012) 6.

170 See generally, R Delgado, 'Words that Wound: A Tort Action for Racial Insults, Epithets, and Name Calling', in *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment,* M Matsuda, C Lawrence, R Delgado, and K Crenshaw (eds), (Westview Press 1993) 89; MJ Matsuda, *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (Westview Press 1993) 89–110.

171 Kukla (n 83) 6 and 16.

assumed – without further argument – that the volume 'The Colonization of Europe. True Speech on Immigration and Islam',[172] which pointed out several incompatibilities between European and Islamic civilisation (e.g., ritual rape), was clearly harmful by invoking misleading arguments. The presumption of harmfulness was based on the tense political situation regarding immigration, and the fact that the book was accessible to the general reader.[173] However, this line of argument is clearly insufficient to prove the existence of actual harm. The *likelihood test* applied in Holocaust denial cases also sets a low burden of proof for the causal link between hate speech and harmful effects. Based on the likelihood test, hate speech is harmful if it is 'directed to inciting or producing imminent lawless action and is likely to incite or produce such action.'[174] Determining the appropriate standard for causal inference is a puzzle that cannot be solved by the reasonable speaker test, for that determination requires a comprehensive theory of probability and causation, a task that goes far beyond the purpose of this paper. Thus, the reasonable speaker test should be coupled with an effective model for causal inference.

Let me conclude this discussion of causal inference with a *caveat*: the claim that certain forms of extreme and anti-democratic speech presuppose an endorsement of militant democracy, precisely as the ECtHR does.[175] The preference for militant democracy, in turn, generally rests on the notion that there is a causal link between hate speech and foreseeable harm to the democratic order, or to the inclusion of a discriminated minority in the legal process. However, if the link between speech and harm cannot be demonstrated – that is, if we cannot be certain that hate speech causally determines harm – then one of the main pillars of the concept of militant democracy is undermined, and the defence of content-based restrictions on freedom of expression weakens. Conversely, the criticism that militant democracy discriminates against 'extreme' or 'anti-democratic' parties and political groups and curtails their fundamental rights gains currency. Thus, the question 'does hate speech

172 *Soulas and Others v France* 15948/03 (ECtHR, 10 July 2008).

173 Ibid paras 36–42.

174 *Brandenburg v Ohio,* 395 US 444, 447 (1969), which partially restates *Schenck* (n 167). *Brandenburg* defends the so-called 'viewpoint neutrality' doctrine: '[C]onstitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of the use of force or of law violation except where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action.'

175 For an extended analysis of 'militant democracy' and its main implication, see the essays contained in A Malcopoulou and AS Kirshner, *Militant Democracy and its Critics: Populism, Parties, Extremism* (Edinburgh University Press 2019).

produce harm?' is closely related to the democratic dilemma 'how much freedom (of speech) should be granted to enemies of democratic values?'. Without sound statistical reasoning, the principle of 'no liberty for the enemy of liberty' could clearly become a pretext for censorship.[176] The issues surrounding the justification of militant democracy – in general, or as a specific approach presupposed by ECtHR jurisprudence – are also beyond the scope of this essay, but certainly represent a limitation of the pragmatic model, at least in all those cases where the European consensus has reached a low level.

7 Tacking Stock

This paper shows that we can, after all, rely on objective, context-sensitive standards for hate speech. Certain linguistic markers and focal points can clearly indicate diminishing language, slurs, prejudices, and stereotypes. These markers are ultimately based on *linguistic conventions* and *pragmatic processes* that take cues from the context of communicative exchange and, ultimately, determine the interpretation space for a particular utterance. I have illustrated how the ECtHR advocates a militant strategy towards hate speech and how the lack of precise definitions paves the way for a test-based approach. Then, drawing on a Gricean account of communication, I distinguished three levels of hate speech. I argued that the identification of hate speech should proceed through a *reasonable speaker test* that establishes situational burdens for plausible deniability. I have explained *inter alia* the main inferential patterns that characterise the reasonable speaker tests, with an eye toward distinguishing between *cancelling* an implicature and *denying* an explicit content. To close the circle, I applied the normative model to several cases decided by the ECtHR and discussed the main limitations of the reasonable speaker tests. The pragmatic approach yields several payoffs.

First, the reasonable speaker test provides a consistent model for dealing with all types of hate speech: a pragmatic support for the multi-pronged or *ad hoc* balancing method generally used by the ECtHR. The model uses deeply rooted linguistic conventions as proxies for communicative intent. More specifically, the reasonable speaker test ascertains the speaker's commitment to hate. When implicatures come into play, the determination of commitment depends largely on fallible calculation of the interpreter involved in the construction of meaning. Commitment attribution is always context-dependent,

176 Heinze (n 7) 129–133.

degree-sensitive, and it may or may not be plausible. It is important to determine the strength of implicit meaning.

Second, as explained earlier, the normative model provides a pragmatic toolbox for evaluating ECtHR cases from a linguistic perspective, and not the only correct answer for every hate speech case or a model for causal inference. There might still be residual hard cases even if the model is applied correctly. For instance, a teacher's mention of 'hordes of Muslims' in France in the school newspaper (which was sanctioned in the *Seurot* case) is borderline.[177] Plausible deniability is always a matter of degree.

Third, the pragmatic account helps to make cases of hate speech 'more epistemologically tractable.' Take misogynistic hate speech, for example. If we assume that an utterance is misogynistic not because it presupposes the presence of a 'constitutive' relationship of patriarchal power or a particular negative feeling on part of the speaker, but because a reasonable speaker would interpret the utterance as misogynistic, we endorse a more objective standard of review.[178] Defining the reasonable speaker in terms of the building blocks of Grice's concept of communication, rather than relying on mere fictions and ideals, provides us with precise protocols and justificatory obligations for identifying the content of utterances based on focal points that underlie human communication. These conventions, in turn, are linked to an analysis of the European consensus on how to understand particular utterances.

Fourth, the model does not constrain additional variables that might be relevant in certain contexts, such as online hate speech. For example, the economic interest of the provider could be considered a relevant characteristic for the appropriate balancing of freedom of expression with competing rights and interests. Judicial balancing and principle-based adjustments are not ruled out.

Author bio

Alessio Sardo (B.A., M.A. in Law Trieste; PhD in Law, Genoa, 2015) is currently Assistant Professor at the University of Genova (Italy) – Law Department. Previously, he was Alexander von Humboldt Fellow at the University of Heidelberg (Germany), Department of Public Law, Constitutional Theory and Philosophy of Law. His research interests include legal theory, comparative constitutional law, general jurisprudence, and human rights. Alessio Sardo's

177 *Seurot v France* 57383/00 (ECtHR, dec, 18 May 2004).
178 K Manne, *Down Girl: The Logic of Misogyny* (Oxford 2018) 60.

published works have appeared on The Modern Law Review, The American Journal of Jurisprudence, The Canadian Journal of Law and Jurisprudence, Archiv für Rechts und Sozialphilosophie, Rechtstheorie, and The Kanazawa Law Journal. Dr. Sardo is also a course convenor and tutor on the Master in Global Rule of Law and Constitutional Democracy, LLM co-run by the Universities of Genoa and Girona, where he teaches Meta-ethics and Constitutional Values. In 2014–2019, he served as a Lecturer in legal philosophy, critical thinking, and general jurisprudence at Bocconi University (Milano). Dr. Sardo was invited as visiting scholar in Kiel, Taiwan, and Paris. He has recently been awarded a Nawa Ulam Grant from the Polish National Agency and a Research Distinction (Honorary Professorship) from the University of Ica (Perù).

List of Recent Publications (Selected):

7) Sardo, A. (2021): ""Border Walls, Pushbacks, and the Prohibition of Collective Expulsions: The Case of N.D. and N.T. v. Spain." Forthcoming in Issue 3 (2021) of the European Journal of Migration and Law.

6) Sardo, A. (2020): "Categories, Balancing, and Fake News: The Jurisprudence of the European Court of Human Rights." The Canadian Journal of Law and Jurisprudence (Issue Aug. 2020 – already published online), pp. 1–26.

5) Sardo, A. (2020): "Robert Alexy, Frederick Schauer, and the 'Positivist' Theses." Archiv für Rechts- und Sozialphilosophie, Beiheft 161, pp. 189–202;

4) Roversi, C., Sardo A. (2019): "Ekins on Groups and Procedures." The American Journal of Jurisprudence 64(1), pp- 79–103;

3) Sardo, A. (2017): "Some Reflections on the Proof-based Theory of Legal Exceptions." The Modern Law Review 80(2), pp. 352–369;

2) Sardo, A. (2017): "Let's Talk about Antinomies: Normative Systems Reloaded." Revus – Journal for Constitutional Theory and Philosophy of Law, pp.1–24;

1) Poggi, F., Sardo, A. (2016): "Do the Right Thing! – Robert Alexy and the Claim to Correctness." Rechtstheorie 47(4), pp.413–441;

AUTHOR QUERIES

NO QUERIES