

Counterspeech

Multidisciplinary Perspectives on
Countering Dangerous Speech

**Edited by Stefanie Ullmann and
Marcus Tomalin**



COUNTERSPEECH

This volume looks at the forms and functions of counterspeech as well as what determines its effectiveness and success from multidisciplinary perspectives. Counterspeech is in line with international human rights and freedom of speech, and it can be a much more powerful tool against dangerous and toxic speech than blocking and censorship.

In the face of online hate speech and disinformation, counterspeech is a tremendously important and timely topic. The book uniquely brings together expertise from a variety of disciplines. It explores linguistic, ethical, and legal aspects of counterspeech; looks at the functions and effectiveness of counterspeech from anthropological, practical, and sociological perspectives; and addresses the question of how we can use modern technological advances to make counterspeech a more instantaneous and efficient option to respond to harmful language online. The greatest benefit of counterspeech lies in the ability to reach bystanders and prevent them from becoming perpetrators themselves. This volume is an excellent opportunity to spread the word about counterspeech, its potential, its importance, and future endeavors.

This anthology is a great resource for scholars and students of linguistics, philosophy of language, media and communication studies, digital humanities, natural language processing, international human rights law, anthropology and sociology, and interdisciplinary research methods. It is also a valuable source of information for practitioners and anyone who wants to speak up against harmful speech.

Stefanie Ullmann is a linguist and postdoctoral research associate at the Centre for Research in the Arts, Social Sciences and Humanities at the University of Cambridge, UK. Her research interests include the use of language in politics and media discourse as well as forms and effects of harmful language in online discourse. She is the author of several journal publications on combatting and mitigating digital harms. Her book *Discourses of the Arab Revolutions* (2022) examines the power and functions of language in sociopolitical conflicts.

Marcus Tomalin has been a member of the Machine Intelligence Laboratory in the Department of Engineering at Cambridge University since 1998. He has published extensively on speech recognition, speech synthesis, machine translation, and dialogue systems, as well as various topics in the philosophy of language and theoretical linguistics, with a recurrent focus on the interconnections between mathematics, logic, and syntactic theory. He has a particular interest in the ethical and social impact of language-based AI systems, and he teaches ethics to undergraduates and postgraduates who are studying philosophy, computer science, and information engineering.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

COUNTERSPEECH

Multidisciplinary Perspectives on
Countering Dangerous Speech

Edited by Stefanie Ullmann and Marcus Tomalin

Designed cover image: Shutterstock Images

First published 2024

by Routledge

4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge

605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2024 selection and editorial matter, Stefanie Ullmann and Marcus Tomalin;
individual chapters, the contributors

The right of Stefanie Ullmann and Marcus Tomalin to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

The Open Access version of this book, available at www.taylorfrancis.com, has been made available under a Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) 4.0 International license. Funded by Knowledge Unlatched.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

Names: Ullmann, Stefanie, editor. | Tomalin, Marcus, editor.

Title: Counterspeech : multidisciplinary perspectives on countering dangerous speech / edited by Stefanie Ullmann, Marcus Tomalin.

Description: Abingdon, Oxon ; New York, NY : Routledge, 2024. |

Includes bibliographical references and index.

Identifiers: LCCN 2023032131 (print) | LCCN 2023032132 (ebook) |

ISBN 9781032454535 (hardback) | ISBN 9781032454504 (paperback) |

ISBN 9781003377078 (ebook)

Subjects: LCSH: Freedom of speech. | Counterspeech. |

Semantics (Philosophy) | Misinformation. | Disinformation.

Classification: LCC JC591 .C68 2024 (print) | LCC JC591 (ebook) |

DDC 323.44/3--dc23/eng/20230905

LC record available at <https://lcn.loc.gov/2023032131>

LC ebook record available at <https://lcn.loc.gov/2023032132>

ISBN: 978-1-032-45453-5 (hbk)

ISBN: 978-1-032-45450-4 (pbk)

ISBN: 978-1-003-37707-8 (ebk)

DOI: 10.4324/9781003377078

Typeset in Sabon

by Taylor & Francis Books



An electronic version of this book is freely available, thanks to the support of libraries working with Knowledge Unlatched (KU). KU is a collaborative initiative designed to make high quality books Open Access for the public good. The Open Access ISBN for this book is 9781003377078. More information about the initiative and links to the Open Access version can be found at www.knowledgeunlatched.org.

CONTENTS

<i>List of illustrations</i>	<i>vii</i>
<i>List of contributors</i>	<i>ix</i>
<i>Acknowledgements</i>	<i>xiv</i>
Introduction <i>Stefanie Ullmann</i>	1
PART I	
Approaches to Counterspeech: Linguistics, Philosophy, and Interdisciplinarity	15
1 Counterspeech Practices in Digital Discourse – An Interactional Approach <i>Sebastian Zollner</i>	17
2 The Philosophy of Counter Language <i>Laura Caponetto and Bianca Cepollaro</i>	50
3 Seeing the Full Picture: The Value of Interdisciplinary Counterspeech Research <i>Joshua Garland and Catherine Buerger</i>	67

PART II

Counterspeech in Context: Media, Culture, and the Legal Framework 83

- 4 Counterspeech as Persuasion and Media Effects 85
Babak Bahador
- 5 Reimagining the Current Regulatory Framework to Online Hate Speech: Why Making Way for Alternative Methods Is Paramount for Free Speech 100
Jacob Mchangama and Natalie Alkiviadou
- 6 Online Hate Speech in Video Games Communities: A Counter Project 123
Susana Costa, Bruno Mendes da Silva and Mirian Tavares

PART III

Automation and the Future of Counterspeech 145

- 7 Automating Counterspeech 147
Marcus Tomalin, James Roy and Shane Weisz
- 8 The Future of Counterspeech: Effective Framing, Targeting, and Evaluation 171
Erin Saltman and Munir Zamir
- 9 Conclusion 196
Marcus Tomalin

- Index 207*

ILLUSTRATIONS

Figures

1.1 N-gram view for the lemma “counterspeech” and variants (Michel et al. 2011).	20
1.2 Comparison of occurrences of “hate speech” and “counterspeech”	22
1.3 Different definitions and semantic distinctions of the concept “counterspeech”/ <i>Gegenrede</i> .	25
1.4 Triadic communication model for hate–counterspeech constructions.	28
1.5 Three dimensions of counterspeech as communicative practice (adapted from Hanks 1996, 230).	30
1.6 Counterspeech reaction by Otto to a racist tweet by Paula (mirror practice).	33
6.1 Screenshot of <i>Divide et Impera</i> .	136
6.2 Screenshot of <i>Youtuber Simulator</i> .	137
6.3 Screenshot of <i>Social Threads</i> .	137
6.4 Screenshot of <i>Deplatforming</i> .	138
6.5 Screenshot of <i>The Invasion of the Cyber trolls</i> .	138
7.1 The encoder–decoder structure of the original transformer model (Vaswani et al. 2017). When predicting the next token, the model attends to the contextual input representations produced by the encoder, along with representations of the previously generated output tokens.	158
9.1 Multimodal counterspeech.	204

Tables

1.1	Transcripts of tweets by Otto and Paula in German and English translation.	34
1.2	Transcript conventions.	34
1.3	Comparison of the most important practices and keywords of the reply levels in comparison.	37
1.4	Fifty keywords for the Otto-Twitter corpus in comparison to the web corpus détente.	38
1.5	Key Discourse Types in Table 1.4.	39
1.6	Transcript of example three of a tweet by User_ComX.	40
1.7	Transcript of counter-conversation “Germans are not only white, blond, blue-eyed people”.	42
6.1	Board 1 – Overview of results of age analysis.	130
6.2	Board 2 – Overview of playing time and exposure to hate speech.	131
6.3	Board 3 – Overview of exposure, victimization, and aggression.	131
6.4	Board 4 – Overview of hate groups.	132
6.5	Board 5 – Overview of online hate speech in everyday life.	132
7.1	An example of a hate speech–counterspeech pair from the MultiCONAN data set.	151
7.2	Automated counterspeech responses to an anti-Semitic prompt.	160
7.3	Results for the DGPT and DGPT-DIA systems for all six metrics.	162
7.4	Results for the DGPT-DIA, GODEL, and ChatGPT systems for all six metrics.	162
7.5	Interrogative percentages for DGPT-DIA, GODEL, and ChatGPT for interrogatives other than “What do you mean?”	163

CONTRIBUTORS

Natalie Alkiviadou is senior research fellow at the Future of Free Speech project. Her research interests lie in the freedom of expression, the far right, hate speech, hate crime, and nondiscrimination. She has published three monographs, namely, *The Far-Right in International and European Law*, *Legal Challenges to the Far-Right: Lessons from England and Wales*, and *The Far-Right in Greece and the Law*. Natalie worked with civil society, educators, and public servants on human rights education and has participated in European actions such as the High-Level Group on Combatting Racism, Xenophobia and Other Forms of Intolerance. Natalie has been the country researcher for the 2019 European Network against Racism report on hate crime and the 2022 report on structural racism. Natalie is an international Fellow (2022/23) of the ISLC – Information Society Law Centre of the Università degli Studi di Milano and a researcher at Global Freedom of Expression Columbia University.

Babak Bahador is research professor at the School of Media and Public Affairs (SMPA) at George Washington University. His research focuses on the overlap of media and politics/international relations, with a primary focus on peace-building. Recent publications include “Monitoring Hate Speech and the Limits of Current Definition”, in *Challenges and Perspectives of Hate Speech Research*, edited by C. Strippel et al. (2023); “Countering Hate Speech Online”, in *The Routledge Companion to Media Misinformation and Populism*, edited by H. Tumber and S. Waisbord (2021); and “Classifying and Identifying the Intensity of Hate Speech”, on the Social Science Research Council *Items* website (2020).

Catherine Buerger is the director of research at the Dangerous Speech Project. She studies the relationship between speech and intergroup violence as well as civil society responses to dangerous and hateful speech online. She is a research

affiliate of the University of Connecticut's Economic and Social Rights Research Group and managing editor of the *Journal of Human Rights*. She holds a Ph.D. in anthropology from the University of Connecticut.

Laura Caponetto is the Sarah Smithson Research Fellow in Philosophy at Newnham College, University of Cambridge. Before joining Newnham, she was a postdoctoral researcher in philosophy at Vita-Salute San Raffaele University and an adjunct lecturer in pragmatics at the University of Pavia. Her research is primarily in applied philosophy of language. She is especially interested in how speech act theory, and pragmatics more broadly, can be brought to bear on issues of moral, political, and societal concern. Her articles have been published in journals such as *Analysis*, *Synthese*, *Topoi*, and *Ethical Theory and Moral Practice*. She is among the founding members of the Italian Society for Women in Philosophy (SWIP Italia). She received her Ph.D. in philosophy from Vita-Salute San Raffaele University in 2019.

Bianca Cepollaro is assistant professor at Vita-Salute San Raffaele University, Faculty of Philosophy. She earned a Ph.D. in philosophy (Institut Jean Nicod, ENS, Paris) and linguistics (Scuola Normale Superiore, Pisa). She was a postdoctoral fellow at IFILNOVA (Lisbon), the University of Milan, and Vita-Salute San Raffaele University. She is interested in social philosophy of language, pragmatics, metaethics, and experimental philosophy. Her research focuses on expressives, toxic speech, and counterspeech, on both theoretical and experimental grounds. Her latest publications include her monograph *Slurs and Thick Terms: When Language Encodes Values* (2020) and papers in journals such as *Philosophy Compass*, *Pacific Philosophical Quarterly*, *Synthese*, and *Linguistics and Philosophy*.

Susana Costa is a Ph.D. student in digital media art at the University of Algarve (UAlg) and the Open University (UAb). She is a collaborator at the Research Center for Arts and Communication (CIAC). Her research has been published in peer-reviewed journals, and she has presented her scientific work at national and international conferences. She has also participated in numerous projects in the field of literacy. As a researcher, Susana focuses on the intersection of education, arts, and technology. Her current research delves into the study of hate speech manifestations and effects in games and online communities of younger people. She proposes art- and game-based approaches to effectively address this pervasive problem. Her most recent publications include "Gamified Installation to Counteract Hate Speech", in *ArtsIT, Interactivity and Game Creation*, edited by A. L. Brooks (2023); "Comparative Study of the Transmedia Element", in *Esports in Esports and the Media Challenges and Expectations in a Multi-Screen Society*, edited by A. Torres-Toukoumidis (2022); and *Video Games and Gamification against Online Hate Speech?* (2021).

Joshua Garland received his Ph.D. in computer science and M.S. in applied mathematics from the University of Colorado. Currently, Joshua serves as the interim director and an associate research professor at Arizona State University's Center on Narratives, Disinformation and Strategic Influence. Previously he served as an Omidyar and Applied Complexity Fellow at the Santa Fe Institute. Dr. Garland combines grounded theory, machine learning, time series analysis, and natural language processing to study a wide variety of complex systems and applications, such as online human social dynamics, disinformation, the climate, ecology, politics, dynamical systems, and more. His recent publications include *Seasonal Temperatures in West Antarctica during the Holocene*, *Collective Moderation of Hate, Toxicity, and Extremity in Online Discussions*, and *Using Scaling-Region Distributions to Select Embedding Parameters*, all published in 2023.

Jacob Mchangama is CEO of the Future of Free Speech Project, research professor at Vanderbilt University, and a senior fellow at the Foundation for Individual Rights and Expression. He is the author of the critically acclaimed book *Free Speech: A History From Socrates to Social Media* and producer and narrator of the podcast *Clear and Present Danger: A History of Free Speech*. He has written and commented extensively on free speech and human rights in international media outlets including *The Economist*, *L.A. Times*, *Washington Post*, BBC, CBS News, NPR, CNN, *Foreign Affairs*, *Foreign Policy*, *The Wall Street Journal*, and *Politico* as well as top-tier academic and peer-reviewed journals. Jacob frequently appears on international TV and radio providing expert commentary on issues related to free speech, tech, and human rights. He is a member of the Danish Free Speech Commission and Forum for Info and Democracy.

Bruno Mendes da Silva is a postdoctoral fellow in Communication, Culture and Arts at the University of Algarve (UAlg) and coordinator professor at the School of Education and Communication (ESEC) of UAlg. He has dedicated the last decade to research on the relations between communication, art, and technology. His most recent publications include “The Enredo Game-Installation: A Proposal to Counter Hate Speech Online in Advances”, in *Design and Digital Communication III*, edited by N. Martins and D. Brandão (2023) and *The Forking Paths: Interactive Film and Media* (2021).

James Roy is a fourth-year engineering student at the University of Cambridge, specializing in information and computer engineering. For his master's thesis, he worked on “Automating Counterspeech in Dialogue Systems” under the supervision of Dr. Marcus Tomalin. During this project, James sought to improve counterspeech generation through the use of reinforcement learning from human feedback by building a data set of human-scored counterspeech responses. Previous research experience includes published work on the effect of neural machine translation on toxicity detection.

Erin Saltman is the director of programming at the Global Internet Forum to Counter Terrorism (GIFCT). She has worked in the technology, NGO, and academic sectors building out counterterrorism strategies and programs internationally. Saltman’s background and expertise include both white supremacy and Islamist extremist processes of radicalization within a range of regional and sociopolitical contexts. Her research and publications have focused on the evolving nature of violent extremism online, youth radicalization, and the evaluation of counterspeech approaches. She was formerly Meta’s head of counterterrorism and dangerous organizations policy across Europe, the Middle East, and Africa. She also spent time as a practitioner working for ISD Global and other CT/CVE NGOs before joining GIFCT. Saltman is a graduate of Columbia University (B.A.) and University College London (M.A. and Ph.D.).

Mirian Tavares is an associate professor in the Faculty of Human and Social Sciences, University of the Algarve. She is the coordinator of the Research Centre in Arts and Communication (CIAC) and director of the Doctorate in Digital Media-Art. She is also vice president of the International Association for Computational Art – ARTECH-Int, a member of the team of curators of the Portuguese Contemporary Art Network, and director of *Rotura – Journal of Communication, Culture, and Arts*. Her most recent areas of research are digital literacy, digital media-art, film, visual arts, and aesthetics. Her recent publications include “IN[The Hate Booth]: A Gamified Installation to Counteract Hate Speech”, in *ArtsIT, Interactivity and Game Creation*, edited by A. L. Brooks (2023); “Street Art, Intersectional Feminism, and Digital Media-Art: Report on the Cyberperformative Artefact ‘Make Me Up!’”, in *Handbook of Research on Urban Tourism, Viral Society, and the Impact of the COVID-19 Pandemic*, edited by P. Andrade and M. Lemos Martins (2022); and *On the Poetics and Analogical/Digital Poetics of José Maçãs De Carvalho* (2021).

Marcus Tomalin has been a member of the Machine Intelligence Laboratory in the Department of Engineering at Cambridge University since 1998. He has published extensively on speech recognition, speech synthesis, machine translation, and dialogue systems, as well as various topics in the philosophy of language and theoretical linguistics, with a recurrent focus on the interconnections between mathematics, logic, and syntactic theory. He has a particular interest in the ethical and social impact of language-based AI systems, and he teaches ethics to undergraduates and postgraduates who are studying philosophy, computer science, and information engineering.

Shane Weisz holds an M.Phil. degree in machine learning and machine intelligence from the University of Cambridge, where he was awarded his degree with distinction after submitting his thesis on “Automating Counterspeech in Dialogue Systems”. Prior to that, he completed his undergraduate studies at the University of Cape Town, where he obtained a bachelor of science (honours)

degree in computer science, mathematics and mathematical statistics. He is currently working as a data scientist at Aerobotics, a South African company that applies AI and machine learning technology to transform the agriculture industry.

Munir Zamir has been immersed in the counterextremism and counterterrorism sector for over 15 years working as a researcher, practitioner, and government consultant. His work has involved practitioner-focused interventions with audiences at risk of radicalization both online and offline. He has also led international strategic communication campaigns within the counternarrative space paired with research aimed at synthesizing the practical considerations of the field with robust conceptual ideas. Mr. Zamir is currently a senior strategist at M&C Saatchi World Services, leading some of their core global partnerships for prevention-focused strategic communication. In parallel, Mr. Zamir is completing his Ph.D. at the University of South Wales, assessing the role of strategic communications within international efforts to counter violent extremism.

Sebastian Zollner is currently a research assistant in German linguistics at the University of Greifswald, specializing in counterspeech and linguistic strategies in social media interactions. In his research, he explores the communicative interventions used to counter discriminatory online content. His ongoing doctoral project on *Counter Speech in Interaction* was recognized with the GAL-Promotion Award (German Society for Applied Linguistics) in September 2022. His research interests include language use in digital media, politolinguistics and framing, intercultural communication, and applied linguistics.

ACKNOWLEDGEMENTS

We want to thank all contributors to this volume for the amazing and important work they do. In addition, we wish to express our gratitude to all of the participants of the workshop “Understanding and Automating Counterspeech”, which took place on 29 September 2021 and which made this volume possible.

For their continued support, we are grateful to everyone at the Centre for Research in the Arts, Social Sciences and Humanities (CRASSH) at the University of Cambridge, especially Steven Connor, Joanna Page, Mette Rokkum Jamasb, and Una Yeung. We are also thankful to Bill Byrne, Ann Copestake, and Ian Roberts.

We are also very grateful to the editors at Routledge, Emily Briggs, Simon Bates, Lakshita Joshi, and Emma Harder-Collins, copy-editor Jennifer Gardner, as well as everyone else involved in the editing and production process of the book. Everyone has been immensely supportive and helpful. Finally, we wish to thank the anonymous reviewers for their valued feedback.

Stefanie Ullmann and Marcus Tomalin

INTRODUCTION

Stefanie Ullmann

The need for counterspeech on- and offline is more pressing than ever before as harmful language and behavior continue to plague interpersonal discourse and communication. Hate speech and (online) hate crimes continue to be on the rise globally. According to multiple studies from the UK, the US, Germany, and Finland, an average of 40% of people stated that they have witnessed harmful content online (Hawdon, Oksanen, and Räsänen 2016; Keipi et al. 2016; Vidgen, Margetts, and Harris 2019). This is particularly concerning given the concrete offline consequences of witnessing and being targeted by hate speech online. Research has confirmed that experiencing hate speech can have traumatic psychological consequences for victims and can spark offline violence (van Geel, Vedder, and Taniilon 2014; Müller and Schwarz 2020). This latter fact goes hand in hand with recent statistics according to which hate crime numbers have surged over the past years in countries like the US and the UK (Carrega and Krishnakumar 2021; OSCE 2021).¹ The police in England and Wales, for instance, have noted a steady increase in the number of hate crimes in the past decade. The most recent numbers show a surge by 26% from 2021 to 2022 (Home Office 2022).

Hate speech commonly affects minorities and the most vulnerable members of society (Office of the United Nations High Commissioner for Human Rights 2021). According to a 2021 report by the Anti-Defamation League (2021), overall, 41% of Americans who took part in the survey had experienced online hate and harassment. More specifically, however, “LGBTQ+ respondents reported higher rates of overall harassment than all other demographics for the third consecutive year” (64%). Jewish participants in the study also reported an increase in experiencing online harassment, from 33% to 36%, compared to the previous year. An even greater increase was observed by African Americans. According to this study, race-related hate and discrimination went up from 42% the previous year to 59% in 2021. In the midst of the COVID-19 pandemic,

the Asian American community “experienced the largest single year-over-year rise in severe online harassment in comparison to other groups”, a surge from 11% to 17%. Overall, the pandemic caused a wave of anti-Asian discrimination, hate speech, disinformation, and hate crimes worldwide (see, e.g., Gover, Harper, and Langton 2020; Costello et al. 2021; Kim and Kesari 2021; Wu, Qian, and Wilkes 2021; He et al. 2022; Han, Riddell, and Piquero 2023).

The spread of online hate and disinformation is one of the biggest challenges and threats democratic societies are facing all over the world right now. According to the Centre for Countering Digital Hate (2023),

Digital technology has changed the way we communicate, build and maintain relationships, set social standards, and negotiate and assert our society’s values. Digital spaces, however, are often safe for bad actors spreading hate and disinformation, turning them into a hostile environment for others.

What Is Harmful or Dangerous Content?

In general, harmful or dangerous content can be defined as any kind of content – be it text, images, or videos – that has the potential to incite violence or aggression against another person or group of people. In the specific case of hate speech, the expressed aggression or hatred is often based on specific characteristics of a person, such as their religion, ethnicity, nationality, sexual orientation, age, disability, etc. However, definitions of hate speech tend to vary, and there is no universally shared understanding of the practice. According to the UN Strategy and Plan of Action, hate speech is defined as

any kind of communication in speech, writing or behaviour, that **attacks** or uses **pejorative** or **discriminatory** language with reference to a person or a group on the basis of **who they are**, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.

(United Nations 2023; emphasis in original)

The international community, however, struggles to define hate speech, particularly in legal contexts, which has become the focus of several research papers (see, e.g., Sellars 2016; Paz, Montero-Díaz, and Moreno-Delgado 2020). Fino (2020, 57) stated that “[d]efining hate speech in international criminal law [...] continues to be elusive”. This lack of a clear definition of harmful content creates challenges for dealing with the problem effectively, as will become clear in the different chapters of this book.

The term “dangerous speech” was coined by Susan Benesch, founder and director of the Dangerous Speech Project. Dangerous speech is defined as “any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group” (Dangerous Speech Project 2023b).

Harmful or dangerous speech, however, does not solely refer to hate speech; it also includes mis- and disinformation. While both terms define the sharing and proliferation of false information, when speaking of disinformation, the spreading of fallacious materials is assumed to be done with intention. Despite being different concepts with different challenges, hate speech and mis- and disinformation are often connected and form complementary parts of a complex problem. When we look at the pandemic, for instance, the proliferation of false information fueled societal polarization. Societies fissured into different ideological groups (e.g., pro- vs. anti-vaccination), some of them drifting toward conspiracy theories. This increasing misunderstanding and miscommunication between people led to a rise in harmful and abusive language, especially on social media platforms (see, e.g., Peters 2022; Tropina 2023). A further challenge to the detection and prevention of harmful content is the increasing use of multimodal elements (e.g., text and video). For example, a sentence like “I love the way you smell” becomes hateful when combined with the image of a skunk (Kiela et al. 2020, 1–2). Text and image may also be joined to establish false connections and become misleading content (Segura-Bedmar and Alonso-Bartolome 2022). These cases make it much harder for humans to discern the true meaning, let alone detection technologies.

The Effects of Hate/Dangerous Speech

Delgado and Stefancic (2009, 363) wrote that “every instance of genocide came on the heels of a wave of hate speech depicting the victims in belittling terms”. From national socialism in Germany to the Rohingya genocide, hate speech against a targeted group of people has always preceded and accompanied these atrocious crimes. In 2021, Facebook was sued for £150 billion by Rohingya members in the US and the UK (Milmo 2021). During the genocide in 2017, the social media platform was used to spread hate speech against the Rohingya community. It was argued that Facebook’s algorithms incited the circulation of genocidal posts. Aziz (2022) wrote that “[w]hile the Facebook authority admitted their failure to moderate hate speech that incited the ethnic violence, the legal case is ongoing and remains an example to keep social media platforms accountable”.

While this case depicts one of the more serious examples of the harmful effects of hate speech, it should be noted that hate speech always affects and harms people. Verbal abuse is a form of violence, and its targets inevitably suffer psychological and potentially even physical consequences. This is supported by numerous research studies that show that direct or face-to-face hate speech causes psychological harm, and victims often suffer from depression, repressed anger, diminished self-concept, and impairment of work or school performances (Delgado and Stefancic 2004, 2009). Moreover, experiencing hate speech online can have offline effects on people, spark violent behavior, and have traumatic consequences for victims (van Geel, Vedder, and Tanilon 2014; Müller and Schwarz 2020). Digital spaces provide a level of anonymity that

seems to lower people's inhibition threshold. Chetty and Alathur (2018, 108) even observed that “[e]xpressing hate speech has become a trend [on online platforms] and people are using this as a shortcut way to get instant popularity without putting more effort”.

As already noted, hate speech most significantly affects and harms minority groups, and hate speech tends to vary depending on the targeted group (Miller Yoder et al. 2022). Moreover, hate speech is becoming increasingly intersectional. One such example is hate speech targeting black women, also known as “misogynoir” (Kwarteng et al. 2022). It has been shown that social networks are nowadays the chief platform for online gender-based harassment (Simons 2015; Chetty and Alathur 2018). Hardaker and McGlashan (2016) identified and studied recurring hate and threats of sexual violence directed at feminist campaigners such as Caroline Criado-Perez. Hate speech also repeatedly targets members of the LGBTQIAP+ community (Ștefăniță and Buf 2021). Empirical research indicates that LGBTQIAP+ people are subjected to high levels of verbal assault, which leads to considerable psychological harm (Nyman and Provozin 2019; Hubbard 2020; Walters et al. 2020). Trans people are even more likely to experience heightened levels of threat and vulnerability and face the lack of support of close ones compared with non-trans LGBTQIAP+ people (Walters et al. 2020). A 2021 report by the Gay and Lesbian Alliance Against Defamation concluded that social media are not a safe space for the LGBTQIAP+ community. In 2022, a study by the Institute for Strategic Dialogue showed that Twitter failed to put a stop to the spread of anti-gay conspiracy theories around Monkeypox, which caused a further spike in homophobic hate speech (Gallagher 2022).

A Norwegian study investigated the effects of hate speech on disabled people (Vedeler, Olsen, and Eriksen 2019). Thirty-eight percent of respondents reported at least one experience of hate speech during the previous 12 months. The authors noted that hate speech “restricts activity and undermines quality of life” and that “people place restrictions on their own lives as a result of being exposed to hate speech, including raising their opinion in public debates” (378). Similarly, a British study (Beadle-Brown et al. 2014) revealed that 46% of disabled people said they had experienced “bad things”.

All in all, groups targeted with hate speech are already often vulnerable to mental health problems and prone to depression, which is only exacerbated by hate and harassment experienced online. One serious consequence is the development of posttraumatic stress symptoms (Dragowski et al. 2011). Other documented effects include a heightened state of fear among victims (Awan and Zempi 2016) and a “silencing effect” (Brown 2015, 198) on its victims, especially members of religious or ethnic minorities. In the context of religious hate speech, Bonotti (2017, 263) wrote that “hate speech which advocates or incites discrimination will undermine the autonomy and self-development of its victims”. Victims are indirectly excluded from democratic self-government and directly affected in that they cannot “enjoy” their basic formal rights and liberties. Finally, Soral, Bilewicz, and Winiewski (2018) found that frequent

exposure to hate speech increases prejudice through desensitization, a problem which will be picked up again in chapter 6 of this volume.

The Need for Counterspeech

The concept of counterspeech is not new, but it has acquired new meaning and importance in the context of digital media and the continuous occurrence of hateful speech and mis- and disinformation. Countering hate and misinformation and advocating for each other and the propagation of truthful information are now timelier and more relevant than ever before. Besides the increase of harmful content on social media platforms, another reason for concern is the content moderation approach taken by the companies. A post or comment may be removed and users can be blocked, if their content violates the company's community guidelines. However, it has been stressed by legal experts that censoring speech can only ever be the last resort (see, e.g., Strossen 2018; Stern 2020; see also chapter 5 in this volume). Instead, in recent years, a counterspeech approach has attracted the attention of researchers and experts from a great variety of different professional and academic fields. "Counterspeech" is understood as "any direct response to hateful or harmful speech which seeks to undermine it" (Dangerous Speech Project 2023a). Benesch and her colleagues have also greatly influenced research on counterspeech (Benesch et al. 2016a, 2016b). Their work has identified and distinguished different types of counterspeech conversations: one-to-one, one-to-many, many-to-one, and many-to-many (Benesch et al. 2016a). Moreover, Benesch et al. devised a taxonomy of eight key counterspeech strategies that are frequently referenced by other scholars and will be discussed further in various chapters of this book. To only briefly mention them here, these strategies include

- 1) presentation of facts to correct misstatements or misperceptions, 2) pointing out hypocrisy or contradictions, 3) warning of possible offline and online consequences of speech, 4) identification with original speaker or target group, 5) denouncing speech as hateful or dangerous, 6) use of visual media, 7) use of humor, and 8) use of a particular tone, e.g. an empathetic one.

(Benesch et al. 2016a, 17)

Research into the real-life application of counterspeech on social media is still in its infancy, but initial results are promising. Studies suggest that counterspeech can be successful; particularly in one-to-one conversations, it has been documented that empathy-based counterspeech can reduce racist hate speech on social media and it has positive effects on bystanders and silent followers of the discussion, decreasing the likelihood of others resorting to harmful language (see, e.g., Benesch et al. 2016a, 2016b; Álvarez-Benjumea and Winter 2018; Garland et al. 2020; Hangartner et al. 2021; Saltman, Kooti, and Vockery 2021).

Counterspeech is in line with international human rights and freedom of speech, and it can be a much more powerful tool against dangerous and toxic speech than blocking and censorship. But we need the knowledge and skills from different experts to understand it better and, thus, use it successfully. This anthology brings together expertise from researchers and specialists from a variety of disciplines to address and tackle the most pressing questions relating to harmful content and the most effective ways of countering it. Specifically, the collection includes contributions from linguists, philosophers, practitioners and civil rights activists, lawyers, media and political communication experts, sociologists, anthropologists, mathematicians, and computer scientists. This volume examines linguistic and philosophical questions such as what constitutes counterspeech, legal considerations of how to assess and respond to potentially harmful digital content, methodological questions as to how to measure and evaluate counterspeech, and even technological deliberations of how automated hate speech detection systems perform at generating counterspeech.

While harmful content has been the focus of many research publications in recent years, the concept of counterspeech has been covered comparatively little. We thus hope to add to closing the gap in the existing literature with this volume. Studies have shown that the greatest benefit of counterspeech lies in the ability to reach bystanders and prevent them from becoming perpetrators themselves. We believe that this book is an excellent opportunity to spread the word about counterspeech, its potential, and future endeavors.

Structure of the Book

The book is structured in three main parts, starting with the forms of counterspeech as well as recommendations on the study of counterspeech from linguistic, philosophical, and interdisciplinary perspectives. The first two chapters of Part I, “Approaches to Counterspeech: Linguistics, Philosophy, and Interdisciplinarity,” offer linguistic and philosophical considerations of counterspeech as a human form of expression. In chapter 1, “Counterspeech Practices in Digital Discourse – An Interactional Approach”, Zollner examines counterspeech from a linguistics perspective. He argues that research on counterspeech so far has been primarily affect-centered and that the linguistic-communicative aspect of it has been neglected. However, counterspeech is in essence a linguistic and communicative practice. Many articles and studies focus on the *what* but not the *how*. Building on already existing approaches from linguistics and rhetoric, such as conversation analysis and narrative and argumentation research, Zollner develops what he terms a “digital interaction linguistics” for the study of counterspeech as linguistic-communicative practice. Zollner also provides an insightful corpus-based analysis of the term “counterspeech” and its German counterpart, *Gegenrede*, and how it has become more frequently used throughout time. He also discusses the term “invectivity” and how it was

coined and defined especially in the context of counterspeech. Zollner exemplifies his arguments with an empirical study of authentic speech data taken from Twitter and shows how the dynamic between hateful speech and counterspeech is interactional and contextual. In his corpus analysis, he shows how, in a real-life online debate between a user and a German company, the original hateful post was used as a template to draft a sophisticated counterresponse. Zollner's study also reveals how multiple users enter a dialogue in response to an initial hateful comment and create counterspeech as "a collective and collaborative practice", merging different argumentative and stylistic strategies.

Chapter 2 continues with a more in-depth discussion of different counterspeech strategies within the context of contemporary philosophy of language approaches. The strategies discussed are denying, blocking, bending, saying nothing, and preemptive counterspeech. On the basis of their discussion, Caponetto and Cepollaro contrive a potential philosophy of counterspeech, or counter-language, as they term it. Their discussion addresses the question of appropriateness and effectiveness of different counterspeech strategies; that is, when they are likely to be successful and when they might fail. Denying, for example, while being the most intuitive response, can also come across as confrontational and may thus fail to effectively counter a toxic statement. The authors further note that hateful or toxic speech may also operate implicitly, such as when common stereotypes are merely alluded to. As an example, Caponetto and Cepollaro use the stereotype that women are worse drivers than men. One does not explicitly have to state that in order for a remark to be sexist (e.g., commenting "Wow! That's huge. No doubt she'll have her husband park it for her" when seeing a woman drive a large SUV). More sophisticated forms of counterspeech, such as blocking, are necessary to object to implicit harmful speech. Another possible strategy is what they call "bending". Caponetto and Cepollaro write that "[b]ending consists in distorting a certain toxic contribution into an innocuous (or at least less toxic) one". It is a clever way of preventing potentially harmful content from entering the "common ground" between speakers in the first place. The toxic statement is immediately stymied.

In chapter 3, Garland and Buerger continue the important discussion of how to best study counterspeech. Specifically, their chapter explores quantitative and qualitative approaches to investigating counterspeech and considers the question of what can be gained by taking an interdisciplinary approach. They note that while both qualitative and quantitative studies of counterspeech exist and are, in fact, complementary, they are rarely performed together. The authors, both with different disciplinary backgrounds themselves, bring together insights from their own research studies, the qualitative case study of the #iamhere network, and the quantitative analysis of Twitter conversations between the two opposing German groups Reconquista Germanica, affiliated with the far right, and Reconquista Internet, actively countering the other group. For the qualitative study, Buerger applied digital ethnographic methods to investigate the practices and motivations of active counterspeakers. This

approach allowed her to gain unique insight into the lives and work of the people who perform counterspeech and the intentions driving their activism. In contrast, Garland and his colleagues quantitatively analyzed tens of millions of tweets, something that is clearly not possible through qualitative means. Specifically, they deployed mathematical analyses to explore the impact and dynamics of the two groups mentioned above. Quantitative approaches of this scale can be of significant importance to improve the detection and classification of both hate and counterspeech. Garland and Buerger make the important point that to study the complexities that underlie human interaction and behavior, the analytical and evaluative tools from different disciplines need to be brought together. Their chapter serves as an important guide to interdisciplinary work and, at the same time, showcases the great importance of this multidisciplinary volume.

Part II, “Counterspeech in Context: Media, Culture, and the Legal Framework”, focuses in more detail on specific questions and issues arising out of the socio-political, cultural and legal contexts in which hate speech and counterspeech exist. In chapter 4, Bahador, on the basis of media and persuasion research, examines how counterspeech influences different audiences. He discusses four distinct audiences – hate groups, violent extremists, the vulnerable, and the public (or bystanders) – and how they each engage with hate and counterspeech. He then continues by examining different counterspeech tactics; that is, how counterspeakers can effectively reach their audiences. These tactics include one-to-one, one-to-many, many-to-one, and many-to-many. After a brief discussion of different methodological approaches to evaluate the effects of counterspeech, Bahador explores the following five effects: the spiral of silence, cognitive dissonance, reframing (rehumanizing), two-step flow, and herding. Hate speech nowadays is mainly an online phenomenon. However, be it off- or online, when people’s views and comments are no longer confirmed, they become less likely to continue making hateful statements. Bahador writes that “[w]hen counterspeakers coordinate their activities in large numbers in [online] spaces, they can change the online norm”. He further identifies reframing as an effective strategy to counter hate speech. Often, hate speech based on nationality or race seeks to dehumanize individuals. By refocusing the narrative on the individual human beings, the collective negative stereotype can be weakened. Moreover, Bahador notes that the presence of counterspeech in online conversations can encourage others, so-called bystanders, to post against hate as well.

In chapter 5, Mchangama and Alkiviadou examine hate speech and counterspeech from a legal and regulatory perspective. As the authors explain in their chapter, currently, social media and big tech companies approach harmful content with removal and blocking, which has sparked great discussions internationally about censorship and freedom of speech. Moreover, the removal of posts relies largely on content moderation, still mostly performed by humans. Content is checked against the company’s internal policy guidelines. In other words, what is considered harmful – and thus qualifies for removal – depends

on each company. Mchangama and Alkiviadou state that “the current regulatory approach needs to be substantially reconsidered and methods including counternarratives must be structurally and systematically advanced”. While free speech does come with harms and costs, “it does not follow that censorship is an appropriate or efficient remedy to combat such harms”. Perpetrators who are banned from a major platform might just relocate to a lesser known site with less strict guidelines. The problem itself, however, remains unaddressed. In this chapter, the authors provide us with an insight into the considerations that feed into the process of policymaking and the development of a regulatory approach to dealing with harmful content. The authors discuss the role of international human rights law as well as the increasing influence of artificial intelligence being used for automated censorship systems.

Chapter 6 takes another empirical and timely approach to hate speech and the potential effects of counterspeech. Costa, Mendes da Silva, and Tavares examine the role of cyberhate in video games and explore ways of countering this threat. Games and the concept of gamification have become an important influence, especially in the lives of young people. While games can be great educational tools, they have also become a “breeding ground for harmful content, racist expression, out-group hatred, online propaganda, sexism, and sexual discrimination”. In this chapter, the authors present the results of a study investigating how young people perceive the exposure to harmful content within the gaming community and discuss possible solutions and counter strategies. A significant correlation, for example, was identified between the hours spent gaming and engagement in harmful behavior. Frequent and repeated exposure tends to desensitize users toward hate speech, which confirms findings of previous studies (e.g., Soral, Bilewicz, and Winiewski 2018). Costa, Mendes da Silva, and Tavares conclude their chapter with a discussion of different gamification scenarios and applications that help raise awareness among users and allow them to engage with the problem of hate speech in games in a meaningful way. The applications require users to engage in social interactions and challenge them to think of ways to minimize hate speech and stop it from spreading.

The final two chapters in Part III of this volume, “Automation and the Future of Counterspeech”, examine the potential for automating the creation of counterspeech and explore important considerations for the future. In chapter 7, Tomalin, Roy, and Weisz explore the task of training dialogue systems to automatically generate counterresponses to harmful speech. The authors outline several scenarios in which an automated counterspeech generator could be useful, such as to assist NGOs in devising counternarratives faster and more efficiently, an approach that was already tested by Chung et al. (2021). First, the chapter discusses different ways of gathering training data: crawling, crowdsourcing, nichesourcing, and hybrid. The authors stress that expert-based multitarget counterspeech data sets are ideal for the training of an automated counterspeech system. Second, the authors address the important task of result

evaluation. Chief properties that need to be evaluated include fluency, toxicity, diversity, and gold-similarity; that is, the response should be as human-like as possible. The chapter then provides an overview of available systems and architectures, before delving into a summary of the results of testing and evaluating current state-of-the-art dialogue systems, such as ChatGPT, for the generation of counterspeech. The authors conclude that none of the systems achieve the best possible scores but that “fine-tuning results can significantly improve counterspeech ability”. Finally, they advise stronger collaboration with NGOs, which can act as human-in-the-loop advisers and ultimately improve counterspeech responses.

In the final chapter of this volume, Saltman and Zamir examine different international best practice examples of counterspeech campaigns. Some of these campaigns developed organically, while others came into existence in cooperation with social media companies. The case studies discussed are: Dutch YouthCan Campaign to Combat Anti-Muslim Biases with Humor, Black Lives Matter, Bring Back Our Girls, Facebook Search Redirect Partnership with Life After Hate, and YouTube Partnership with Abdullah X. The authors’ work is rooted in counterextremism and counterterrorism. In the first section of their chapter, Saltman and Zamir discuss framing strategies and the questions of who you are trying to reach and what you want to achieve with counterspeech. They point out that counterspeech nowadays needs to accommodate the multimodal nature of communication, both online and offline. The authors appropriately mention that as forms of extremist and violent content diversify, especially on digital platforms, counterspeech must evolve accordingly. Another section examines counterspeech campaigns in more detail and guides the reader through their chosen examples. The following section tackles the complex yet important question of how to evaluate and measure the effectiveness of such campaigns. They note the importance of knowing one’s audience and performing sentiment analysis to “understand what type of knowledge, awareness, and behavioural change has or has not occurred”. Saltman and Zamir conclude with considerations for the future of counterspeech, emphasizing that approaches must be more dynamic and, for example, adopt a more “fluid” understanding of audiences as active agents. Equally, the diversity of platforms needs more attention, and counterspeech strategies must be devised and adapted across platforms.

All in all, this book offers a varied blend of theoretical foundations, the development of new ideas and approaches, as well as insightful case studies of counterspeech campaigns, empirical analyses of authentic social media data, and experimental explorations of gamification and automation. The key insights will be summarized again in a conclusion at the end of the volume.

A final note on spelling conventions. Different spellings exist for counterspeech, such as counter speech, counter-speech, and Counterspeech. For the sake of conformity, we have opted for the lowercase version written as one word. Moreover,

throughout the book, you will encounter different terms being used instead of “hate speech”, including “dangerous speech”, “toxic speech”, and “cyberhate”. They all represent variants of what is essentially harmful speech.

Note

- 1 An ordinary crime becomes a hate crime when a perpetrator chooses a victim because of a specific characteristic such as their nationality or sexual orientation.

References

- Anti-Defamation League. 2021. *Online Hate and Harassment: The American Experience 2021*. 22 March 2021. <https://www.adl.org/online-hate-2021>. Accessed 1 June 2023.
- Álvarez-Benjumea, Amalia, and Fabian Winter. 2018. “Normative Change and Culture of Hate: An Experiment in Online Environments.” *European Sociological Review* 34 (3): 223–237. doi:10.1093/esr/jcy005
- Awan, Imran, and Irene Zempi. 2016. “The Affinity between Online and Offline Anti-Muslim Hate Crime: Dynamics and Impacts.” *Aggression and Violent Behavior* 27: 1–8.
- Aziz, Abdul. 2022. “Rohingya Diaspora Online: Mapping the Spaces of Visibility, Resistance and Transnational Identity on Social Media.” *New Media & Society*. doi:10.1177/14614448221132241
- Beadle-Brown, Julie, Lisa Richardson, Colin Guest, Aida Malovic, Jill Bradshaw, and Julian Himmerich. 2014. *Living in Fear: Better Outcomes for People with Learning Disabilities and Autism*. Main Research Report. Canterbury: Tizard Centre, University of Kent.
- Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016a. *Counterspeech on Twitter: A Field Study*. A Report for Public Safety Canada under the Kanishka Project. <https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/>
- Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016b. *Considerations for Successful Counterspeech. Evaluating Methods to Diminish Expressions of Hatred and Extremism Online*. Kanishka Project of Public Safety Canada. <https://dangerousspeech.org/considerations-for-successful-counterspeech/>
- Bonotti, Matteo. 2017. “Religion, Hate Speech and Non-Domination.” *Ethnicities* 17 (2): 259–274. doi:10.1177/1468796817692626
- Brown, Alexander. 2015. *Hate Speech Law: A Philosophical Examination*. New York: Routledge.
- Carrega, Christina, and Priya Krishnakumar. 2021. “Hate Crime Reports in US Surge to the Highest Level in 12 Years, FBI Says.” CNN. 26 October 2021. <https://edition.cnn.com/2021/08/30/us/fbi-report-hate-crimes-rose-2020/index.html>. Accessed 4 September 2023.
- Centre for Countering Digital Hate. 2023. *About*. <https://counterhate.com/about/>. Accessed 7 June 2023.
- Chetty, Naganna, and Sreejith Alathur. 2018. “Hate Speech Review in the Context of Online Social Networks.” *Aggression and Violent Behavior* 40: 108–118. doi:10.1016/j.avb.2018.05.003
- Chung, Yi-Ling, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021. “Empowering NGOs in Countering Online Hate Messages. Online Social Networks and Media.” *Online Social Networks and Media* 24. doi:10.1016/j.osnem.2021.100150

- Costello, Matthew, Long Cheng, Feng Luo, Hongxin Hu, Song Liao, Nishant Vishwamitra, Mingqi Li, and Ebuka Okpala. 2021. "COVID-19: A Pandemic of Anti-Asian Cyberhate." *Journal of Hate Studies* 17 (1): 108–118. doi:10.33972/jhs.198
- Dangerous Speech Project. 2023a. *Counterspeech*. <https://dangerousspeech.org/counterspeech/>. Accessed 5 September 2023.
- Dangerous Speech Project. 2023b. *What Is Dangerous Speech?* <https://dangerousspeech.org/about-dangerous-speech/>. Accessed 5 September 2023.
- Delgado, Richard, and Jean Stefancic. 2004. *Understanding Words That Wound*. Boulder, CO: Westview Press.
- Delgado, Richard, and Jean Stefancic. 2009. "Four Observations about Hate Speech." *Wake Forest Law Review* 44 (2): 353–370.
- Dragowski, Eliza A, Perry N Halkitis, Arnold H Grossman, and Anthony R D'Augelli. 2011. "Sexual Orientation Victimization and Posttraumatic Stress Symptoms among Lesbian, Gay, and Bisexual Youth." *Journal of Gay & Lesbian Social Services* 23 (2): 226–249. doi:10.1080/10538720.2010.541028
- Fino, Audrey. 2020. "Defining Hate Speech: A Seemingly Elusive Task." *Journal of International Criminal Justice* 18 (1): 31–57. doi:10.1093/jicj/mqaa023
- Gallagher, Aoife. 2022. "Monkeypox and 'Groomers': How Twitter Facilitated a Hate-Riddled Public Health Disinformation Campaign." Institute for Strategic Dialogue. 11 August 2022. https://www.isdglobal.org/digital_dispatches/monkeypox-and-groomers-how-twitter-facilitated-a-hate-riddled-public-health-disinformation-campaign/. Accessed 4 September 2023.
- Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. "Countering Hate on Social Media: Large Scale Classification of Hate and Counter Speech." In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 102–112. Association for Computational Linguistics. doi:10.18653/v1/2020.alw-1.13
- Gay and Lesbian Alliance Against Defamation. 2021. *2021 Social Media Safety Index*. <https://glaad.org/smsi-2021/>. Accessed 1 June 2023.
- Gover, Angela R, Shannon B Harper, and Lynn Langton. 2020. "Anti-Asian Hate Crime during the COVID-19 Pandemic: Exploring the Reproduction of Inequality." *American Journal of Criminal Justice* 45: 647–667. doi:10.1007/s12103-020-09545-1
- Han, Sungil, Jordan R Riddell, and Alex R Piquero. 2023. "Anti-Asian American Hate Crimes Spike during the Early Stages of the COVID-19 Pandemic." *Journal of Interpersonal Violence*, 38 (3–4): 3513–3533. doi:10.1177/08862605221107056
- Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, and Karsten Donnay. 2021. "Empathy-Based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment." *Proceedings of the National Academy of Sciences* 118 (50). doi:10.1073/pnas.2116310118
- Hardaker, Claire, and Mark McGlashan. 2016. "Real Men Don't Hate Women: Twitter Rape Threats and Group Identity." *Journal of Pragmatics* 91: 80–93. doi:10.1016/j.pragma.2015.11.005
- Hawdon, James, Atte Oksanen, and Pekka Räsänen. 2016. "Exposure to Online Hate in Four Nations: A Cross-National Consideration." *Deviant Behavior* 38 (3): 254–266. doi:10.1080/01639625.2016.1196985
- He, Bing, Caleb Ziem, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2022. "Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis." In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21)*, 90–94. New York: Association for Computing Machinery. doi:10.1145/3487351.3488324

- Home Office. 2022. *Hate Crime, England and Wales, 2021 to 2022*. 6 October 2022. <https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2021-to-2022/hate-crime-england-and-wales-2021-to-2022>. Accessed 1 June 2023.
- Hubbard, Luke. 2020. *Online Hate Crime Report: Challenging Online Homophobia, Biphobia and Transphobia*. London: Galop. <https://galop.org.uk/wp-content/uploads/2021/06/Galop-Hate-Crime-Report-2021-1.pdf>
- Keipi, Teo, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2016. *Online Hate and Hateful Content: Cross-National Perspectives*. New York: Routledge.
- Kiela, Douwe, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine. 2020. “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.” doi:10.48550/arXiv.2005.04790
- Kim, Jae Yeon, and Aniket Kesari. 2021. “Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech during the COVID-19 Pandemic.” *Journal of Online Trust and Safety* 1 (1). doi:10.54501/jots.v1i1.13
- Kwarteng, Joseph, Serena Coppolino Perfumi, Tracie Farrell, Aisling Third, and Miriam Fernandez. 2022. “Misogynoir: Challenges in Detecting Intersectional Hate.” *Social Network Analysis and Mining* 12 (166): 1–15. doi:10.1007/s13278-022-00993-7
- Miller Yoder, Michael, Lynnette Hui Xian Ng, David West Brown, and Kathleen M Carley. 2022. “How Hate Speech Varies by Target Identity: A Computational Analysis.” In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 27–39. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Milmo, Dan. 2021. “Rohingya Sue Facebook for £150Bn over Myanmar Genocide.” *The Guardian*, 6 December 2021. <https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence>. Accessed 1 June 2023.
- Müller, Karsten, and Carlo Schwarz. 2020. “Fanning the Flames of Hate: Social Media and Hate Crime.” *Journal of the European Economic Association* 19 (4): 2131–2167. doi:10.1093/jeea/jvaa045
- Nyman, Hanna, and Annastasiya Provozin. 2019. “The Harmful Effects of Online and Offline Anti LGBTI Hate Speech.” Master’s Thesis, Linnæus University, Sweden. <http://www.diva-portal.org/smash/get/diva2:1355445/FULLTEXT02>
- Office of the United Nations High Commissioner for Human Rights. 2021. *Report: Online Hate Increasing against Minorities, Says Expert*. 23 March 2021. <https://www.ohchr.org/en/stories/2021/03/report-online-hate-increasing-against-minorities-says-expert>. Accessed 1 June 2023.
- OSCE. 2021. *Hate Crime Data*. <https://hatecrime.osce.org/hate-crime-data?year=2021>. Accessed 4 September 2023.
- Paz, María Antonia, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. “Hate Speech: A Systematized Review.” *SAGE Open* 10 (4). doi:10.1177/2158244020973022
- Peters, Michael A. 2022. “Limiting the Capacity for Hate: Hate Speech, Hate Groups and the Philosophy of Hate.” *Educational Philosophy and Theory* 54 (14): 2325–2330. doi:10.1080/00131857.2020.1802818
- Saltman, Erin, Farshad Kooti, and Karly Vockery. 2021. “New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis.” *Studies in Conflict & Terrorism* 46 (9): 1547–1574. doi:10.1080/1057610X.2021.1888404
- Segura-Bedmar, Isabel, and Santiago Alonso-Bartolome. 2022. “Multimodal Fake News Detection.” *Information* 13 (6): 284. doi:10.3390/info13060284

- Sellars, Andrew. 2016. "Defining Hate Speech." Berkman Klein Center Research Publication No. 2016–2020, Boston Univ. School of Law, Public Law Research Paper No. 16–48. doi:10.2139/ssrn.2882244
- Simons, Rachel Noelle. 2015. "Addressing Gender-Based Harassment in Social Media: A Call to Action." In *iConference 2015 Proceedings*, 1–6. Urbana: University of Illinois. <https://hdl.handle.net/2142/73743>
- Soral, Wiktor, Michał Bilewicz, and Mikołaj Winiewski. 2018. "Exposure to Hate Speech Increases Prejudice through Desensitization." *Aggressive Behavior* 44 (2): 136–146. doi:10.1002/ab.21737
- Ștefăniță, Oana, and Diana-Maria Buf. 2021. "Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research." *Romanian Journal of Communication and Public Relations* 23 (1): 47–55. doi:10.21018/rjcp.2021.1.322
- Stern, Kenneth S. 2020. *The Conflict over the Conflict: The Israel/Palestine Campus Debate*. Toronto: New Jewish Press.
- Strossen, Nadine. 2018. *Hate: Why We Should Resist It with Free Speech, Not Censorship*. New York: Oxford UP.
- Tropina, Tatiana. 2023. "Pandemics and Infodemics: How COVID-19 Is Reshaping Content Regulation?" In *Beyond the Pandemic? Exploring the Impact of COVID-19 on Telecommunications and the Internet*, edited by Jason Whalley, Volker Stocker, and William Lehr, 229–243. Bingley: Emerald. doi:10.1108/978-1-80262-049-820231011
- United Nations. 2023. *Understanding Hate Speech*. <https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech>. Accessed 5 September 2023.
- van Geel, Mitch, Paul Vedder, and Jenny Tanilon. 2014. "Relationship between Peer Victimization, Cyberbullying, and Suicide in Children and Adolescents: A Meta-Analysis." *JAMA Pediatrics* 168 (5): 435–442. doi:10.1001/jamapediatrics.2013.4143
- Vedeler, Janikke Solstad, Terje Olsen, and John Eriksen. 2019. "Hate Speech Harms: A Social Justice Discussion of Disabled Norwegians' Experiences." *Disability & Society* 34 (3): 368–383. doi:10.1080/09687599.2018.1515723
- Vidgen, Bertie, Helen Margetts, and Alex Harris. 2019. *How Much Online Abuse Is There? A Systematic Review of Evidence for the UK*. Policy Briefing. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf. Accessed 7 June 2023.
- Walters, Mark A, Jennifer Paterson, Rupert Brown, and Liz McDonnell. 2020. "Hate Crimes against Trans People: Assessing Emotions, Behaviors, and Attitudes Toward Criminal Justice Agencies." *Journal of Interpersonal Violence* 35 (21–22): 4583–4613. doi:10.1177/0886260517715026
- Wu, Cary, Yue Qian, and Rima Wilkes. 2021. "Anti-Asian Discrimination and the Asian-White Mental Health Gap during COVID-19." *Ethnic and Racial Studies* 44 (5): 819–835. <https://doi.org/10.1080/01419870.2020.1851739>

PART I

Approaches to Counterspeech: Linguistics, Philosophy, and Interdisciplinarity



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

1

COUNTERSPEECH PRACTICES IN DIGITAL DISCOURSE – AN INTERACTIONAL APPROACH

Sebastian Zollner

Counterspeech Research from a Linguistic Perspective

Counterspeech research can look back on a relatively young research history and has not yet been more comprehensively addressed and analyzed from a linguistic perspective. Overall, Bahador (2021, 510) stated for the research landscape that counterspeech is still “under-researched, undertheorised and underdeveloped”. In the following, I present some selected aspects of counterspeech research to finally show which potentials can be tapped by a linguistic approach.

There are several reasons for the lack of linguistic research in particular. So far, the focus has been on researching hate speech and digital violence practices and less on the linguistic-communicative examination of counterspeech. While numerous efforts have been made in linguistics to (methodologically) explore hate speech (Meibauer 2013; Klinker, Scharloth, and Szczek 2018) that make an important contribution to nuancing, defining, and educating about the phenomenon, counterspeech is usually only touched upon and addressed in passing (e.g., Baider and Kopytowska 2018; Geyer, Bick, and Kleene 2022).

To be sure, linguistic studies also acknowledge the need for counterspeech, as when Hoffmann (2020, 47) concluded with regard to invective acts such as racial discrimination and other forms of hate speech, “Where there is no perceptible counterspeech, the boundaries are pushed further”. Meibauer also pointed out that “presuppositional protest” (Meibauer 2022, 24) can be an important means of countering derogatory speech acts, which has also been described elsewhere as “blocking” (see also Langton 2018). However, a systematic description of the means by which this happens or can happen has yet to be provided.

Another reason for this is what is seen as characteristic of counterspeech: Language is often seen only as a vehicle for content-related messages (e.g., of tolerance and solidarity) or as a thematic counter-presentation of facts and

apparently does not represent the central feature for describing counterspeech. Accordingly, there are many content-analytic approaches to studying counterspeech, looking for “content that challenges extremism online” (Bartlett and Krasodomski-Jones 2015) or attempting to classify social media comments as hate speech or counterspeech via machine learning and annotated training data (Chung et al. 2019; see also chapter 7 in this volume). This is usually done using the counterspeech taxonomy, which was conducted by the Dangerous Speech Project in their pioneering studies (Benesch et al. 2016)¹:

1. Presenting facts to correct misstatements or misperceptions
2. Pointing out hypocrisy or contradictions
3. Warning of offline or online consequences
4. Affiliation
5. Denouncing hateful or dangerous speech
6. Visual communication
7. Humor
8. Tone

It is noted that these strategies can also be combined with each other – for example, in a tweet or a post – which is because the strategies are also on different levels and cannot be disjunctively delimited. They mainly describe *what* the counterspeakers do and which contents and modalities are used. The *how* – that is, the linguistic-communicative structure of counterspeech practices – is neglected, or without interactional embedding and without context-sensitive methodological description. Thus, Wright et al. concluded that counterspeech acts are not produced in a pattern-like manner and that no linguistic-stylistic commonalities can be discerned:

Counterspeech acts can assume many forms. Crucially, in our review of known counterspeech acts, we have observed no indication that these forms are templated - meaning that any two arbitrary counterspeech acts will not share language, syntax, or style.

(Wright et al. 2017, 58)

The question “How to do things with counterspeech?”, following speech act theorist John Austin’s 1955 Harvard lectures “How to Do Things with Words” (Austin 1962), has thus not yet been clarified and represents an open desideratum, which also primarily concerns linguistics. Therefore, I argue for a stronger focus on the linguistic means of counterspeech to develop a differentiated repertoire of elements and constituents of counterspeech, which can ultimately also serve as a basis for further machine learning methods; for example, by elaborating certain linguistic indicators for counterspeech and using them in the future for counterspeech taxonomies and also as machine learning classifiers (in addition to possible classifiers such as hashtags, user profiles, social media platforms; cf. Wright et al. 2017).

The fact that nothing is known about the linguistic characteristics of counterspeech apart from content, addressee tailoring, and the category “tone” is due, on the one hand, to the research and cognitive interests outlined above, which are predominantly very effect centered, to the research methods and to the nature of the data basis or how the counterspeech data examined were collected or were created. While many studies and counterspeech data are prototypical, introspective, or elicit in nature, this chapter argues for investigating counterspeech empirically, focusing on authentic speech data (cf. the case study Otto vs. Paula as discussed in this chapter) that were not produced for experiments or communication science simulations and evaluations. Examples of such approaches include the experiment on normative changes in users through counterspeech comments in Álvarez-Benjumea and Winter (2018), the self-authored counterspeech in Miškolci, Kováčová, and Rigová (2020) or the examples from the CONAN (COunter NArratives through Nichesourcing) data set (Chung et al. 2019), in which experts were specifically acquired to produce counternarratives.²

Despite the (banal) realization that counterspeech is mainly linguistic (whether oral or written) communication, the linguistic-communicative side of the phenomenon has been neglected, especially by linguistics and related disciplines, although this opens up an important and exciting field of work, especially for applied linguistics. Thus, as indicated, the previous landscape of counterspeech research has not been interested in linguistic features of counterspeech and has rather focused on its topics or contents as well as its effectiveness or merits compared to censorship measures in terms of freedom of speech. Furthermore, it makes sense to apply already existing approaches from linguistics and rhetoric – for example, from conversation analysis as well as narrative and argumentation research – to the subject of counterspeech, such as positioning strategies, contradiction techniques, and persuasion attempts. In the following, I will try to develop a digital interaction linguistics that can be used to perform micro-analyses of hate-counterspeech dynamics, focusing on both the linguistic structures of counterspeech, discursive strategies, and propositions. The aim of this chapter is to contribute to the identification of some identity criteria for counterspeech practices, taking into account both linguistic means and sequential embeddedness, to establish a theory of counterspeech as a linguistic-communicative practice. Indications for such identity criteria can be found, for example, in conflict linguistics, which deals with how conflicts are linguistically carried out and processed (especially dialogical forms of conflict such as arguing).

The necessity of an empirical and interactional discussion can also be deduced from the conceptual history of the terms “counterspeech” and *Gegenrede* used in English and German. Therefore, in the following paragraph I would like to elaborate the interactional core of counterspeech.

A Brief History of Terms and Usage of Counterspeech/*Gegenrede*

What is the meaning of counterspeech (intension)? To which phenomena can I refer with these terms (extension)? In the two subsections “The Lemma ‘Counterspeech’ in English” and “The Lemmas ‘Counterspeech’ and *Gegenrede* in German”, I illuminate the history of the terms and the emergence of the terms counterspeech, as far as I could determine with the query in different German- and English-language corpora.

In the third subsection, “Meanings of Counterspeech”, the different readings and the different conceptual ranges of counterspeech and its German counterpart *Gegenrede* are brought together and traced. It will be shown that the meaning spectrum of the lexeme contains not only an obviously reactive moment but also an interactive and proactive reading. The recent public discourse on the concept of counterspeech is also reflected in the semantic enrichment of the lexeme. The English term counterspeech as well as the literal German translation *Gegenrede* are polysemous expressions; that is, words that have several readings. Depending on the context, one or the other partial meaning is activated.

I argue from the history of terms and meanings that an interactional approach can do justice to the object of *Gegenrede* in the first place.

The Lemma “Counterspeech” in English

If one looks up the lemma counterspeech or its variants in the relevant dictionaries of English, no current entry can be found for the noun or the verb “to counter-speak,” although the expression is increasingly used in current discourses (see Figure 1.1). As word formation constituents, the words are composed of the prefix “counter-” and the noun “speech” or the verb “to speak.” The prefix counter- is productive and it can occur in combination with “almost any substantive expressing action, as motion, counter-motion, current, counter-current, or even to any word in which action or incidence is imputed, as measure, counter-measure, poison, counter-poison” (*Oxford English Dictionary* [OED] 2023a). All *verba*

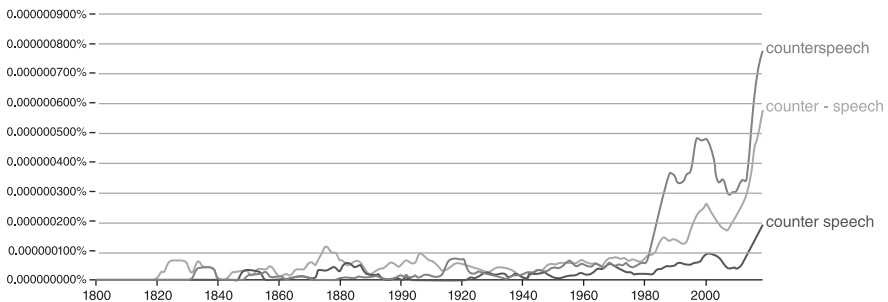


FIGURE 1.1 N-gram view for the lemma “counterspeech” and variants (Michel et al. 2011).

dicendi can also be prefigured with counter-, such as counter-arguing, counter-narrative, counter-propaganda. These are listed in the OED as regular “permanent compounds” with the prefix counter-, counterspeech does not belong to this group. By means of the OED, the semantic reading can be specified in more detail, which also applies to counterspeech: “Done, directed, or acting against, in opposition to, as a rejoinder or reply to another thing of the same kind already made or in existence” (OED 2023a). The OED, although listing the lemma “counterspeech” with the meaning “[s]peaking against, contradiction” has labeled the entry “obsolete”, which “usually means that no evidence for the term can be found in modern English” (OED 2023b). There is linguistic evidence for the given meaning from 1647, and the entry was last updated in June 2021.

In the GoogleBooks corpora of English, it can be seen that there is evidence in modern English for the lexeme counterspeech (and its orthographic variants), all of which experienced an initial boom from 1980 onwards and have again been found significantly more frequently in the GoogleBooks corpus since the 2010s (see Figure 1.1). The first rise in the 1980s parallels the emergence of the concept of “hate speech”, whose coinage is attributed to Mari Matsuda and critical race theory in the late 1980s (Matsuda 1989). In the course of this, the term counterspeech is increasingly used in the legal literature in the United States in relation to so-called hate group speech, so-called fighting words, obscenity cases, and the legal interpretation of the 1st Amendment (e.g., Kalven 1960; Downs 1985; Sunstein 1986).

In the linguistically processed corpora of English, the Corpus of Historical American English, there is no evidence for counterspeech, and in the Corpus of Contemporary American English there are nine hits for the variant counterspeech, all from a scholarly article by Nadine Strossen (1997).

In the 2.1-billion-word Oxford English Corpus (a text corpus of 21st-century English) there are also examples where the lemma counterspeech is found in media discourse and legal discourse, with almost all of the evidence coming from the subcorpus to American English, which is another indication that the discourse on this, at least under this catchphrase, is mainly in the United States. This brief diachronic examination of the lexeme counterspeech in English shows that the word is not widely used in everyday or professional discourse outside of legal discourse. Its low-frequency usage makes it easier to enrich the word with new shades of meaning in discourse and to conceptualize it more strongly, which can also be observed at the latest from the 2010s by various practitioners from the field of extremism prevention, activists, NGOs, and think tanks (see Bahador 2021) but also companies such as Facebook (today’s Meta).³ As Wright et al. (2017) noted, different counterspeech concepts with different scopes can be distinguished: a narrow one that defines only immediate responses as counterspeech, a medium scope definition that encompasses any content that counters or contradicts hateful or extremist content, and a very broad understanding of counterspeech that goes beyond the actual hate–counterspeech discourse and encompasses jurisprudence, forms of education, and public events for informational and entertainment purposes.

Finally, we provide a brief comparison with the term hate speech and its occurrences in the GoogleBooks corpus. The progression curve in the n-gram viewer clearly shows the increasing and overall more frequent use of the expression hate speech in contrast to its “antagonist” counterspeech. In 2018, hate speech was about 26.5 times more frequent than any form of counterspeech, with this factor increasing to about 38.8 in 2019 (see Figure 1.2). This quantitative preponderance of hate speech could be interpreted as an indication that both academic research and public debate are placing a greater focus on regulating hate speech rather than on penalties to promote counterspeech.

The Lemmas “Counterspeech” and *Gegenrede* in German

In the German-speaking world, the English term counterspeech is often literally translated as *Gegenrede*. However, the two terms are not easily equated, as I will show in the following.

Compared to English, the term *Gegenrede* is much more frequent in German and is also listed in all important dictionaries and reference works. Especially in German, the polysemy of the term becomes clear, which, however, does not (yet) capture the discourse specificity of the expression in the context of “hate on the net”.

The *Digital Dictionary of the German Language* (2023) gives two readings for the lemma *Gegenrede*: (1) response and (2) rebuttal, contradiction.

Gegenrede in the first reading means a simple answer, which is temporally and sequentially subordinated to another utterance. Here, the alternating change of speech and counterspeech as a dialogical principle is in the foreground. This understanding of *Gegenrede* originates from the understanding of the ancient philosophical dialogue, which of course is very different from the “messy dialogue” of everyday life:

Thinking about dialogue, one may also be reminded of one of dialogue’s oldest paradigmatic written forms: philosophical dialogue. Philosophical

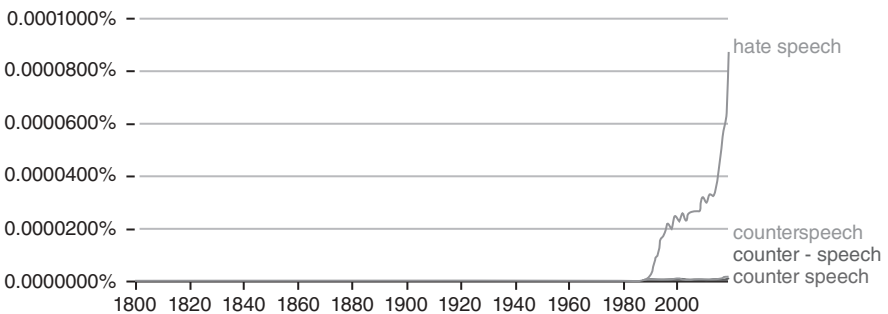


FIGURE 1.2 Comparison of occurrences of “hate speech” and “counterspeech”

dialogue is defined as a verbally conducted argument involving two or more persons. Their interaction consists in speech and counter-speech, which can take the shape of questions and answers (to clarify terms and concepts), claims and rejections (to offer judgement and evaluation), and proof and falsification (to reach a conclusion) (see Lorenz 2005: 189). Even though philosophical dialogue purports to report actual discussions between, for example, Socrates and a number of dialogue partners it is clear that their representation in writing is rhetorical and literary in quality (Hösle 2006) and has very little resemblance to the “messy” forms of dialogic exchange one can find in actual conversations.

(Mildorf and Thomas 2017, 3)

A compellingly contradictory quality is given to *Gegenrede* only in the second reading as *Widerrede* or *Widerspruch*. A *Gegenrede* thus presupposes an antecedent that is contradicted with the reactive utterance or to which the *Gegenrede* (in content and language) places itself in a contradictory relation. In this second sense, *Gegenreden* are conceivable against all possible representations of facts with which one does not agree. Counterspeech that refers to a preceding discourse contribution can therefore be produced by anyone, and actors from right-wing populist or anti-democratic spectrums also claim this for themselves.⁴ In the self-perception of such actors, their own statements, declared as hate speech by others, are nothing more than legitimate speech, contradiction, disagreement, or at best harsh criticism. Or as Konstanze Marx summed it up, “All these interpretations converge roughly in the meaning of ‘unwanted *Gegenrede* not fit for public consumption’” (Marx 2020, 708, translated from German to English by the author).

If one wants to use counterspeech as an analytical concept, this aspect makes it clear that, in addition to the spatiotemporal and content-related relation to an antecedent, a moral-normative dimension must also be included in the practice of counterspeech, which is also named as a characteristic of counterspeech practices in various definitions from research; for example, with “messages of tolerance” (Rieger, Schmitt, and Frischlich 2018, 464), “prosocial behavior” (Marx 2020, 708, translated from German to English by the author; see also Marx and Zollner 2020), or the enforcement of an “antidiscriminatory stance” (Fedtke and Wiedemann 2020, 96).

A query in the written language part of the DeReKo (German Reference Corpus) yields almost 5000 hits for the lemma *Gegenrede* since the 18th century, with the majority occurring in the years after 2002. Looking at the chronologically sorted records, it becomes clear that *Gegenrede* is mainly used for speeches and counterspeeches in political debates (in parliaments as well as in television broadcasts) until the late 1990s. *Gegenrede* is an integral part of political debate and democratic communication. *Gegenrede* can thus be assigned as a vocabulary of institutional political lexis; for instance, when it comes to replicating government statements. There are no specific topics in

which *Gegenrede* is used particularly frequently. *Gegenrede* is also a genre of the media; for example, when well-known journalists such as Rudolf Augstein use *Gegenrede* or when talk shows use *Gegenrede* to question politicians on a topic. Particularly as a result of the increasing establishment of social media, *Gegenrede* appears to be gaining a stronger foothold in discourse, especially against such phenomena. In 2010, journalist Patrick Gensing stated: “Gegenrede is in demand here. Because right-wing radicals and neo-Nazis are also active in the virtual world” (Gensing 2010).

Since 2015, the broader reception of the concept of counterspeech, which was adopted from the legal discourse in the United States (see above), has led to further semantic enrichment, which is not yet reflected lexicographically in dictionaries. The first evidence for the use of counterspeech in German-language publications in the sense presented here I found in the 2000 jurisprudential doctoral thesis on border demarcations by Elisabeth Holzleithner. In 2007, Daniel Thürer mentioned in his newspaper article “Racism as Poison” the possibility of combating racism and Holocaust denial with counterspeech in a democratic society (Thürer 2007). The third oldest evidence comes from the publication of jugendschutz.net (2013), which is already relevant to the anti-hate speech discourse, titled *Rechtsextremismus Online Beobachten und Nachhaltig Bekämpfen* (*Watching right-wing extremism online and fighting it sustainably*). It was not until 2015 that evidence for the lemma counterspeech can be found in relevant corpora such as the Digital Dictionary of the German Language WebXL corpus or DeReKo. The reasons for this can be seen in the increasing political efforts worldwide (see Gagliardone et al. 2015) to effectively combat hate speech. The activities of Facebook, which founded a counterspeech initiative in 2015 with the support of the then-German Justice Minister Heiko Maas, were particularly influential in shaping the discourse.

Counterspeech, borrowed from English, thus functions in German-language discourse as an antonym to hate speech; that is, this partial meaning is immediately apparent in German-language contexts but has also semantically enriched the German-language word *Gegenrede*. I show this process in the next section.

Meanings of Counterspeech

The antecedent to which one reacts with counterspeech is thus fixed to “evil”, in our case hate speech, dangerous speech, or extremist propaganda or invective acts of communication as an overarching phenomenon (for an explanation of invectivity, see section Analyzing Invectivity). Counterspeech selects these preceding acts to counter them in an act of contradiction, whose linguistic-communicative form and proposition thus enter into an oppositional relation to the antecedent, which can be accessed and made visible through the analysis of authentic empirical data.

Warnke (2021) critically noted that a reduction of *Gegenrede* to a reactive dimension with a dyadic basic communication structure (A reacts to B) is not

adequate to the subject matter and that this is a semantic narrowing of *Gegenrede* that does not cover the entire spectrum of *Gegenrede* possibilities. With reference to Strossen (2018), he explained that counterspeech can also be used as advocacy. This changes the direction in which counterspeech is addressed: away from focusing on the emitters of hate toward solidarity with individuals or groups affected by hate speech and their public defense. This public advocacy is then no longer done with a focus on an antecedent but against the background of explicitly stated or presupposed values (such as the equality and dignity of all people, tolerance, respect) and thus contrasts the preceding hate speech with something positive. Counterspeech with these proactive dimensions is structured triadically: someone speaks to someone for the benefit of someone else (see Warnke 2021). Counterspeech responds relationally to a trigger event (the antecedent), which requires a context-sensitive analysis that takes into account how *Gegenrede* is embedded in larger communicative projects and activities and through which linguistic-communicative as well as interactive means this relation is established, so that these are also recognizable for all those involved in the interaction. Counterspeech in the modeled understanding, as shown by the findings from the corpora, as well as the discussion of previous scientific research, contains a moral-normative dimension, which also emerges through indicators on the linguistic surface in interaction with the contextual factors.

Nadine Strossen's view on counterspeech corresponds to the second reading opposition to antecedent, but she emphasizes that counterspeech is not a means that must necessarily confront hate speech but that counterspeech is the central procedure that can be used on any topic circulating in the "marketplace of ideas":

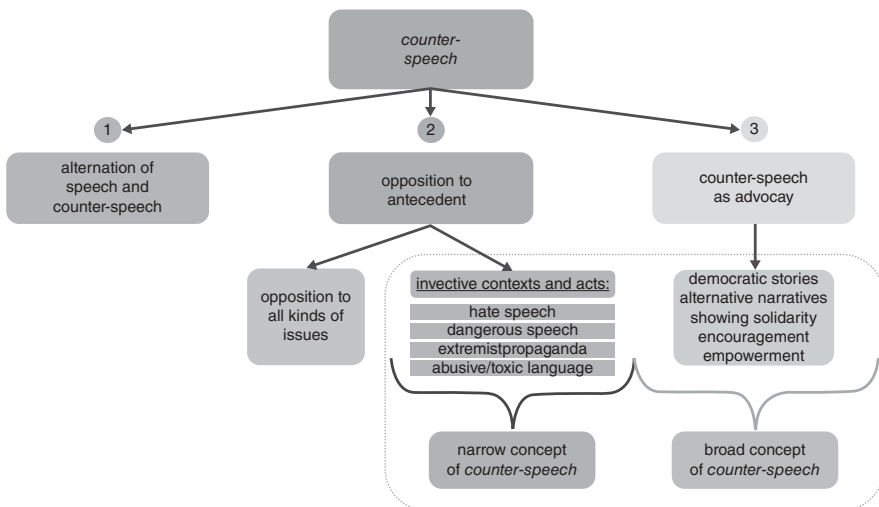


FIGURE 1.3 Different definitions and semantic distinctions of the concept “counter-speech”/*Gegenrede*.

- “The term ‘counterspeech’ encompasses any speech that counters a message with which one disagrees” (Strossen 2018, 158).
- “This term refers to any speech that counters or responds to speech with a message that the speaker rejects, including ‘hate speech’” (Strossen 2018, xxii).

From a legal perspective, which would like to restrict freedom of speech as little as possible, this broad definition may be appropriate, but as an analytical category, to say something about the specific practices of counterspeech, in my opinion it is necessary to apply some further criteria. I would like to present these criteria in the following section in a working definition.

Counterspeech in Interaction and Invectivity

At this point, I would like to present a definition of counterspeech that takes into account the previous explanations and attempts to establish counterspeech as a theoretically and methodologically grounded category of analysis.

Counterspeech is a linguistic-communicative practice that responds directly and immediately or spatiotemporally to invective practices, such as discriminatory events, derogatory and hurtful speech, or symbolically or linguistically mediated ideologies of inequality, or, in a broader sense, proactively advocates for those affected by invective practices. Content, linguistic form, purpose, and addressees depend on the concrete invective constellation (nature of the trigger event/antecedent, arena, medium, actors) and the communicative activity carried out. The content, form, and function of counterspeech practices are in a relation of contradiction/opposition to their respective antecedent, which they thereby mark as an invective trigger event.

This perspective on counterspeech emphasizes the importance of viewing counterspeech not as an isolated textual product but as an interactive activity co-constructed between two or more interactants. Rather than limiting ourselves to the status quo of a particular textual artefact (is it counterspeech or not?), we should conceive of counterspeech as a process that also unfolds in the digital realm via successive social media posts (“turns”). It is also possible to observe multiple secondary turns in response to an initial post.

Analyzing Invectivity

In this section, I would like to introduce the concept of “invectivity”, which was developed in the special research area 1285 of the same name, located at the TU Dresden, for the description of hate–counterspeech dynamics and counterspeech practices in particular. The authors of the conceptual group understand invective as follows:

Invectivity is meant to bring into focus those aspects of communication (verbal or non-verbal, oral, written, gestural or pictorial) that are apt to

disparage, harm or exclude. However, the manifestations and functions of invective discourse – understood as a mode of performing invectivity – do not conform to a rigid pattern, but rather appear in complex, historically variable constellations in the media, politics, in social situations and the aesthetic sphere.

(Ellerbrock et al. 2017a)

The new term or category of analysis functions as an umbrella term for a series of invective phenomena and thus brings into focus the commonalities and differences as well as the specific historical, social, and medial conditions of certain invective practices, which can thus never be considered without their context and especially not without their follow-up communication. For the question of whether a given social media comment is, for example, hate speech – that is, an invective practice – or not, it is crucial how it is reacted to in the follow-up comments. Consequently, invective practices can also be imagined on a scale between more or less invective, depending on how they are received: For the variable constellations of invectivity can “appropriately be understood only as performative events, as a mesh of attributions, resonances, and follow-up communications, and in the context of their social, discursive, and medial enabling conditions” (Ellerbrock et al. 2017b, 4). As a consequence of this programmatic observation, a methodological procedure results that takes into account this interactional network of connection communication. In the following I will show two ways to unravel and describe this network (one is a corpus-linguistic approach and the other a micro-discourse–interactional approach).

Invectives act in the name of the “whole people”; groups of people feel invested as part of “the whole nation” or a religious community; third parties intervene vicariously with regard to the perceived violation of the integrity of minorities and marginalized groups, etc. Constellation analysis thus ultimately pursues a reconstruction of the dynamics of invective networks and the relational self- and other-positionings possible within them.

(Ellerbrock et al. 2017b, 13)

Metainvectivity and Counterspeech

Metainvective reflexivity refers to communication processes that clarify the invective (insulting or hurtful) content of utterances, classify them within a normative interpretive framework, and reflect on their possible effects. This concept is often applied to examine the boundaries of invective, often in the context of a change in media, such as from face-to-face conversation to social media or to different art forms. Whether an utterance is classified as legitimate criticism, a joke, or an insult or violence depends on many factors, including the discursive meaning of certain words and phrases, the expressive rights of the

participants, the context of the invective, and the political implications involved. Metainvective reflexivity looks at one or more of these parameters and examines the implications and geo-social relevance of invective. It provides reflexive accessibility to the moral orders that are used for legitimation and makes them observable and analyzable for researchers. In this context, metainvective utterances can themselves have an invective potential (cf. SFB Invektivität 2021).

Counterspeech in Invective Constellations

The invective triad is an ideal-typical model for analyzing invectivity that considers three key roles: the offending party (invective), the offended party (in-vecive), and a testifying or legitimizing instance (for example, audience or gee-whiz). This dynamic constellation underlines the importance of the context and circumstances of an invective utterance and emphasizes that such utterances are always embedded in a web of enactments and perceptions of the actors involved (see Figure 1.4). The positions of actors in this triad are diverse and can change dynamically, with role changes and discourse-strategic recoding typical. The invective triad serves as a

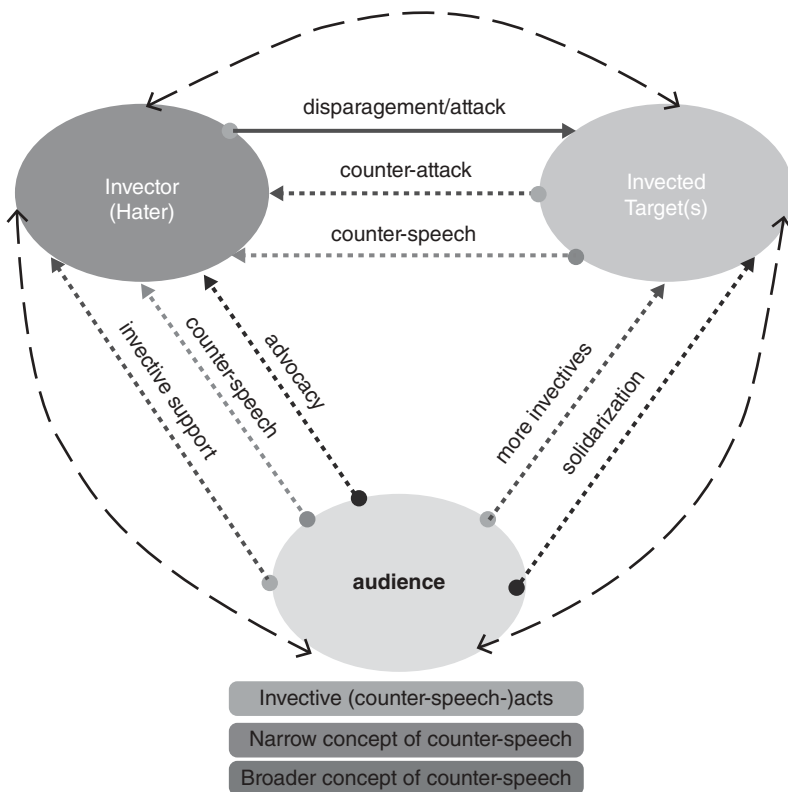


FIGURE 1.4 Triadic communication model for hate-counterspeech constructions.

starting point for analyzing the complexity, dynamics, and practice forms of invective, rather than promoting a homogeneous or linear conception of the phenomenon. It also provides an entry point to the power relations expressed in positioning and exposure within invective.

All processes of interaction and communication can take place in the mode of the invective, for which a wide variety of invective practices are employed in different spatiotemporal arenas, media, and social constellations, which then break through into concrete invective situations. Hate speech, incitement, cyberbullying, shit-storms, and other analog and digital communication phenomena can be subsumed under this perspective and focused under a common theoretical and methodological angle. Invective communication follows a triadic basic structure consisting of the invectives (e.g., bodyshamers), the invected (e.g., those affected by bodyshaming), and an audience (bystanders, silent readers).

The research program on invectivity goes well beyond a traditional reading of the speech-act model of verbal abuse. Invective occurs under conditions that involve a complex interweaving of utterances and the attribution of meaning to such utterances in the reactions of addressees and observers.

(*Ellerbrock et al. 2017b, 8*)

In Butler's terms, *Gegenrede* is only possible because the effects of invective and hurtful speech acts on those addressed are not finally determined. This opens up "the possibility of a counterspeech, a kind of speaking back," which represents an "alternative to the endless search for legal remedies" (Butler 2006, 30–31). This opening "space" can now be used for *Gegenrede*, or counterspeech. The opening of this space is thus a result of entering into interaction to manage the perceived invective (e.g., a sexist joke, an anti-Semitic presupposition, etc.) in a situation. *Gegenrede*, therefore, I would like to conceive of as a tool, a linguistic-communicative practice of invective management that can take on a wide variety of forms and functions. How counterspeech can be conceptualized as a practice is presented in the following section.

Counterspeech as a Practice of Invectivity Management

In the definition above as well as in the title, I refer to counterspeech practice and counterspeech practices. This classification goes back to considerations that originate in sociological practice theory and are also productively used in linguistics in the following and adapted for its own purposes.

Thus, counterspeech can be described in its function as a communicative practice for the purpose of invective management. At this point, I use a concept of practice with several levels based on Hanks (1996) as well as the linguistic-communicative determinants of practices as described by Deppermann, Feilke, and Linke (2016) (see Figure 1.5). Hanks developed a model for the characterization of communicative practices that comprises three dimensions: the

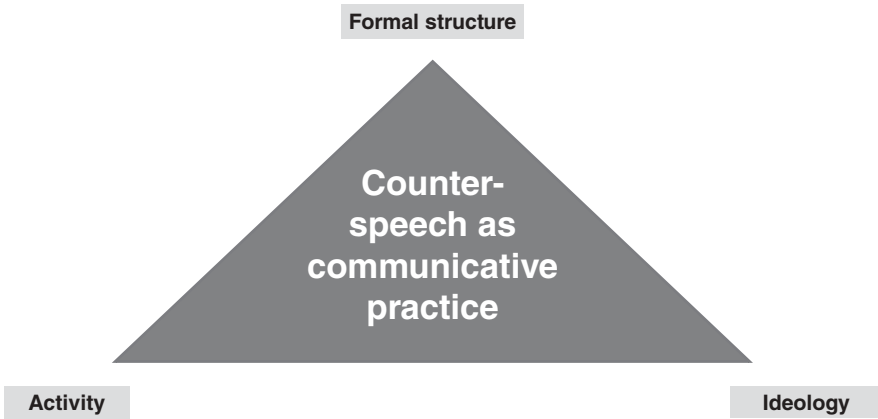


FIGURE 1.5 Three dimensions of counterspeech as communicative practice (adapted from Hanks 1996, 230).

linguistic form and structure, the assignable (linguistic) actions or partial practices, and the ideologies and perspectivizations that can be found in them by the actors involved.

Examining *Gegenrede* as a linguistic-communicative practice in interaction means assuming that not every counter-speaker in a given invective constellation has to “invent” individual actions from scratch to respond to an invective comment. To solve the extremely complex task of, for example, countering a sexist posting, the speaker can fall back on certain linguistic-communicative building blocks from his or her individual practice household that have been acquired beforehand during the accomplishment of certain tasks or in didactic settings.

This also means that these practices do not have a uniform formal language, so that one could automatically draw conclusions about the practice via phenomena on the linguistic surface (for example, certain words or grammatical phenomena). Bick, Geyer, and Kleene (2021, 102) demonstrated that the nominal phrase “evil Muslims,” embedded in a construction like “the oh-so-evil Muslims are to blame,” can be considered as part of an ironic counterspeech practice. Through the ironic reversal, stereotypical and anti-Muslim argumentation patterns of the invective reference utterance can be exposed. This makes it clear that despite the negative attributions on the surface of the language, it can still constitute a form of counterspeech.

Digital counterspeech practices, like all linguistic-communicative practices, evolve in relation to social structures of purpose and are subject to historical change (Deppermann, Feilke, and Linke 2016). For many people, reacting to racist remarks and other variants of verbalized group-related misanthropy is certainly not an everyday communicative task for which they can fall back on proven routines. An analog precursor of digital counterspeech is the countering

of so-called regulars' table slogans, for which Klaus Peter Hufer developed a special argumentation training in the 2000s that addresses the need to position oneself routinely and argumentatively against right-wing populist and discriminatory statements (see Hufer 2016). However, the conversational solutions practiced there have not yet been extensively studied linguistically. In a conversation-analytic study, Kuck (2019) elaborated on argumentative strategies and pointed out insinuations to debunk racist thinking. Counterspeech practices for the purpose of invective management can also be vividly described with the metaphor of communicative work, in which move by move, speech contribution by speech contribution is interactively and/or collaboratively worked on certain common or opposing goals. I have described various discursive varieties of this invective management via counterspeech elsewhere. These include positioning work, correction work, blocking work, irritation work, activation work, defense work, escalation/de-escalation work, empathy and understanding work, and solidarity work (see Zollner 2022).

Accordingly, the description of the form of the linguistic-communicative (and social) phenomenon that is the focus of the investigation is an essential part of the determination of a communicative practice. This is done with an interactional linguistic perspective on hate-opposition dynamics and attempts to describe how these specific interactional dynamics of these social interactions are produced with linguistic resources and which (pattern-like/pattern-forming) linguistic means are used for the purpose of invective management (cf. Selting 2016). This functional determination is reflected in certain linguistic activities (communicative activities) or partial practices of counterspeech, which can serve this purpose in a specific situation, because "practices are characterized by the context-sensitive use of certain linguistic-communicative forms as resources for solving basic tasks of interaction constitution and for producing certain handlings" (Deppermann, Feilke, and Linke 2016, 1). The performed partial practices/actions, in a given linguistic-structural form in each case, are always also shaped by the ideological (*weltanschauliche*) positioning of the actors involved. In the concrete instances of in situ practices, these three dimensions, according to Hanks (1996), always play inseparably into each other.

For counterspeech research, it is of central importance to determine the unit/measure of *Gegenrede*. So far, *Gegenrede* has often been studied as a responsive one-turn act to the reply and recorded which action or speech act is performed in relation to the antecedent:

In Hanks' (1996, 242) view, in order to analyze communicative practices, it is first fundamental to identify a unit of description "that is greater than the single utterance but less than a language."

(Lanwer and Coussios 2018, 134)

With Wolfgang Imo (cf. Imo 2016), I would like to suggest characterizing counterspeech as a phenomenon of medium scope: Counterspeech itself is not

an action, but it can be built up by quite different actions such as counter-arguing, ironizing, doubting, or questioning. The practice of counterspeech thus consists of a set of one or more actions to solve a communicative task (e.g., countering racism) and to achieve a goal (“invective management”). The respective concrete practice can be carried out by a singular action/partial practice or by linking several actions. This means that singular postings, each with a single action/subpractice, can be understood as a counterspeech practice but also that more complex counterspeech practices can be described, such as what I would like to call counter-conversations. Counterspeech could then be understood as a secondary practice, composed of a combination of primary acts or subpractices to accomplish a particular communicative task and achieve the goal of invective management.

This approach takes the *doing* of counterspeech seriously and is interested in its linguistic manifestation, the so-called counteriness. This emerges through the established relation and the interactive processes that make the counter-orientation visible. Trying to describe counterspeech as a practice and doing counterspeech means the following: The term “practice” in this case emphasizes that counterspeakers do something with the “linguistic units and signals they use” and that the addressed invectives and the audience understand them in terms of what they do (Selting 2016, 29). Practices are thus constituted through “recurrent use and combination of particular resources or bundles of resources [...] for the production and contextualization of a particular action in a particular sequential context” (Selting 2016, 29, translated from German to English by the author), which will be illustrated in the following analysis chapter. By showing how participants in counterspeech interactions make clear “what is going on here”.

Analysis and Corpus-Linguistic Introduction

In the following, I present different approaches to present counterspeech as a practice of invective management. First, I present a corpus linguistic approach to the Twitter (today’s X) case study Otto vs. Paula, which complements the ethnomethodological data collection procedure and participant observation. This interaction secures the analysis of the concrete invective constellation (see Figure 1.6). This is followed by an interactional-linguistic and microanalytic sequence analysis of so-called counter-conversations, in which the resource bundles of *Gegenrede* practices are then targeted. Both ways allow us to identify patterns that can be typical for counterspeech practices.

Paula’s racist tweet (1) also contains a screenshot of Otto’s homepage, which shows a black advertising model, as well as a screenshot of a complaint mail to Otto in which the rejection of this advertising practice is justified and Otto is to be made to justify this practice by asking questions (see Table 1.1 for the transcripts of tweets and Table 1.2 for the transcript conventions). Paula posts the post late at night and the tweet initially circulates only in the circle of her right-wing followers, because user Paula is a self-confessed Alternative for Germany

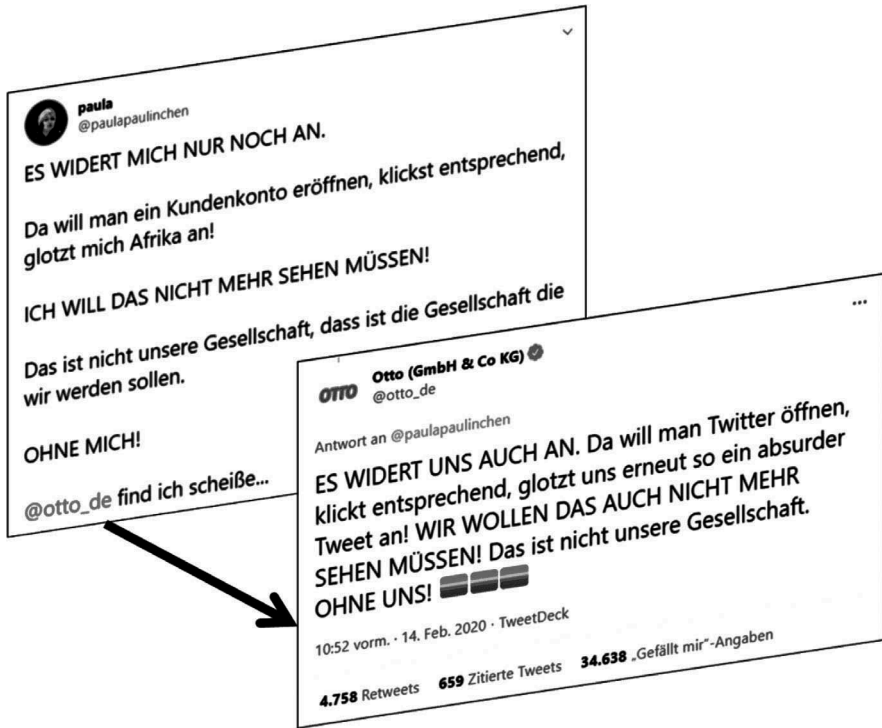


FIGURE 1.6 Counterspeech reaction by Otto to a racist tweet by Paula (mirror practice).

voter and engages in right-wing propaganda with her activist account. Otto's reply (2) receives wide coverage and thus the original tweet also goes viral. At the time of data collection, 4378 Twitter replies could be collected under this conversation, which also constitutes the corpus for this analysis.

The Otto Twitter corpus contains 79,386 types or 120,462 tokens. There are 2636 different Twitter accounts participating in the discourse, although 2037 of them post only once. In addition, there are a few power users in this discourse, who have contributed between 50 and 76 times in different places in the Twitter reps. It is certainly no coincidence that the top three accounts with the most tweets are among those defending the racist initial tweet, spreading right-wing populist slogans and trolling other users (cf. Meier-Vieracker 2020).

Template-Based Counterspeech

First, the initial tweet of Paula Paulinchen was analyzed with regard to thematic and linguistic-communicative composition, according to the thesis of Housley et al. (2017) that most counter-comments refer to, revisit, verbalize, refute, or contrast the themes or linguistic structures set in the initial tweet. In analyzing

TABLE 1.1 Transcripts of tweets by Otto and Paula in German and English translation.

(Nr.)	<i>Tweet Text</i>	<i>Metadata (& Media)</i>
(1)	<p>ES WIDERT MICH NUR NOCH AN. Da will man ein Kundenkonto eröffnen, klickst entsprechend, glotzt mich Afrika an! ICH WILL DAS NICHT MEHR SEHEN MÜSSEN! Das ist nicht unsere Gesellschaft, dass ist die Gesellschaft die wir werden sollen. OHNE MICH! @otto_de find ich scheiße...</p> <p>IT JUST DISGUSTS ME. You want to open a customer account, click accordingly, and Africa stares at me! I DON'T WANT TO HAVE TO SEE THIS ANYMORE! This is not our society, this is the society we're meant to become. WITHOUT ME! @otto_de I think sucks....</p>	<p>Paula (TW-OTTO-0001) 🕒13.02.20 22:55:16 🔄219/👍673/🗨️892</p>
(2)	<p>↳ @Paula ES WIDERT UNS AUCH AN. Da will man Twitter öffnen, klickt entsprechend, glotzt uns erneut so ein absurder Tweet an! WIR WOLLEN DAS AUCH NICHT MEHR SEHEN MÜSSEN! Das ist nicht unsere Gesellschaft. OHNE UNS! 🚩🚩🚩</p> <p>@Paula IT DISGUSTS US TOO. You want to open Twitter, click accordingly, and again such an absurd tweet stares at us! WE DON'T WANT TO HAVE TO SEE THIS ANYMORE EITHER! This is not our society. WITHOUT US! 🚩🚩🚩</p>	<p>Otto GmbH (TW-OTTO-0001) 🕒14.02.20 09:52:16 🔄4754/👍34633/🗨️1034</p>

TABLE 1.2 Transcript conventions.

(1)	<i>Reference number in body text</i>	↳	<i>Reply to previous comment (new/next level)</i>
Meta-data	<p>Otto GmbH [Username/Pseudonym] (TW-OTTO-0001) [Corpus ID] 🕒 Timestamp of comment 🔄 Retweets or shares/👍Likes/🗨️ Comments</p>	<p>🔍 or (2)🔍</p>	<p>Reply to the same comment to which the previous comment replies (same level as the previous comment) or Reply to reference (no.)</p>
		+	Multiple connected tweets

the initial tweet, the goal was to identify interactional features that anticipate or significantly shape subsequent responses and thus delineate the corridor of counterspeech.

The initial tweet introduces the discussion by beginning with an expressive speech act in capital letters, “It just disgusts me,” offering the reader insight into the author’s emotional stance. She provides a rationale via a brief chronological narrative of “opening a customer account,” which begins with her visit to Otto’s homepage and ends with her feeling “gawked at by Africa.” This

implicitly suggests that black people are automatically considered Africans and should exclusively appear as advertising material in this context. This implicature will be reconstructed and discussed in the following articles. The pronoun *that* in “don’t want to have to see that anymore” refers both anaphorically to what was mentioned before and cataphorically to the attached screenshot of Otto’s homepage. This screenshot shows a black model advertising a new customer promotion with a 5% discount, which facilitates the interpretation of the implicature. At the same time, the author uses this modal construction to contextualize her inevitability in the process and the necessity of looking at it. Overall, this linguistic design reveals an ideological positioning against the representation of diversity in advertising on Otto’s homepage, which leads to a subsequent discussion and counterspeech.

The author of the tweet makes a distanced evaluation of this situation by claiming that this is not “our society”. By using the personal pronoun, she assigns herself to an ingroup and positions the model – and implicitly all those who resemble it – outside of society, in an outgroup that is not supposed to belong to society. Her following “Ohne mich!” (“Without me!”), written in capital letters, can be interpreted as an announcement of a boycott.

However, “Without me” is ambiguous in its positioned in-betweenness and does not refer exclusively to her role as a customer of Otto. It can also be understood as a dissociation against the processes of social change she proclaims. In this context, the author conveys a clear ideological positioning toward the representation of diversity and social change by signaling a categorical rejection and distancing herself from these processes.

Finally, Otto_de is addressed via the @-operator, in which the advertising slogan “Otto ... find ich gut” is bastardized and instead evaluated with “finde ich scheiße”.

In addition to the homepage screenshot, there is a screenshot of an email that makes this clear. In the email, she refers to black people and people of color only as “colored people”, “people of African descent”, “Africans”, and contrasts these terms with white “German consumers”, showing a world “where people of color are always present”. With this posting, Paula Paulinchen follows up on her previous Twitter activities in which she collects BIPOCS in German-language advertising to support her observations and critique. The point of reference is a passage from the UN migration pact, which supposedly states that BIPOCs must be represented more strongly in the media.

Against the background of her profile history, her racist criticism can be interpreted as an offer to her followers to collaboratively establish a racist evaluation of economic advertising practices, also because racism is expressed here in a rather implicit and coded way. With Durrheim, Greener, and Whitehead (2015) it can be concluded that a communicative project (Linell 2012) is started here, in which racism is to be further elaborated collaboratively. Since Paula does not yet use the racism markers of the strongest category, there is room and connection possibilities for trumping practices or a competition in

which the racist comments cascade. At the beginning and later, collaborating “dialogic partners” appear again and again, supporting Paula’s theses or relativizing and legitimizing what Paula meant. An evaluation of the top users with the most tweets belonging to this Conversation_Id/Thread shows that especially the supporters send off many tweets to defend content and person. This distorts the picture of opinions in the comment column and should show the potential audience that there is strong and numerous support among the followers of the right-wing bubble (cf. Giles 2021).

With the *totum pro parte* “glotzt mich Afrika an”, Paula opens up the space for supposedly ontologically based semantic contiguities such as “African – black skin color” and thus activates a membership categorization device that is taken up in various reply instances, whereby different membership categories are used and contrasted with each other. However, many users explicitly point out that the invocation of this ontological contiguity is not contemporary or inaccurate and that being black and being German have not been mutually exclusive for a long time. These phenomena and their associated sequences are examples of counterspeech as invective management.

As I will show, this structure of the tweet and its blueprint is the foil against which many counterspeech practices are performed. In doing so, the tweet makes thematic specifications such as Otto’s being a customer, the composition and change of society, the lexical field of disgust, practices of corporate evaluation (customer feedback/complaints), and group affiliations. For this purpose, the tweet activates the “membership categorization device”, here, for example, with “inhabitants of continents”, “skin color”, but also exclusive pro-forms I/us/we that polarize. Paula’s tweet thus sets up parameters for the follow-up tweets in terms of content and topic but also in terms of linguistic function, as Housley et al. (2018) also stated for Twitter campaigns. This pre-structuring opens up many points of contact that can be used as a starting point for acts of contradiction. It is quickly obvious that Otto also uses the trigger tweet as a template in his reaction, but what is interesting are the variations and the now different deictic reference structures, which are now to be resolved differently. Otto’s viral reply tweet now uses the initial tweet as a template for his counterspeech practice, as Housley already noted.

As Giles (2021) noted, it is often the case on Twitter that as a tweet goes viral, controversy and contradiction increase. Again, this is the case where the tweet first circulates among Paula’s circle of followers and earns approval. It then receives greater attention primarily through Otto’s response, which should become the company’s most successful tweet in 2020. In this way, the tweet leaves its original filter bubble, so to speak.

To complement the ethnographic-qualitative approach, a corpus-linguistic approach is also suitable for a first access to concrete hate-counterspeech dynamics – for example, via keyword or frequency analyses – to validate the individual observations if necessary.

Table 1.3 shows that two discourse spaces have differentiated in the follow-up communication: on the one hand, direct replies to Paula and, on the other hand, direct replies to Otto. The replies to Paula are clearly more critical, antagonistic, and invective (especially toward Paula) than the tweets that can be found under the Otto reply, which clearly refer positively to Otto's reaction, announce support, and comment metadiscursively on the conversation.

If we zoom in one level higher on the entire corpus and filter out stop words, we find the following keywords, which were calculated in comparison with the web corpus detente (see Table 1.4).

The top 50 keywords in Table 1.4 indicate that this is a discourse column dominated by counterspeech (see Table 1.5 for the key to the color codes used). The keywords have high keyness values, which means that they are very typical for the Otto corpus compared to the reference corpus. They were also checked in their respective sequential context to ensure that the majority of them were used in the sense presented here. Thus, via this view, one can get a first impression of which topics and discourse strands are dominant in hate-counterspeech dynamics. From there, one can then zoom in to individual points in these dynamics and look at how these themes are interactively negotiated. It becomes clear that there are some

TABLE 1.3 Comparison of the most important practices and keywords of the reply levels in comparison.

<i>Subcorpus: Replies to Paula</i> →Majority opposition		<i>Subcorpus: Replies to Otto</i> →Majority praise and encouragement	
Indignation and co-contradiction		Evaluation of Otto's Reply Tweet	
Connection to political Discourses		Meta-discursive comments (<i>block, report</i>)	
Management of Membership Categories relevant to racism		Supportive and commissive speech acts towards Otto but also towards the Black model as well as the Black community as a whole.	
Invective counter-attacks		Positioning practices for diversity	
"Mirror"-Constructions			
Single Keywords	Bi-Grams	Single Keywords	Bi-Grams
<i>du</i>	<i>Anderer Hautfarbe</i>	<i>Ehre</i>	<i>Klares Wort</i>
<i>damit</i>	<i>Liebe Paula</i>	<i>Geblockt</i>	<i>Gute Reaktion</i>
<i>widerst</i>	<i>Braune Scheiße</i>	<i>Beste</i>	<i>Klasse Reaktion</i>
<i>Maul</i>	<i>Hey Paula</i>	<i>Klasse</i>	<i>Beste Antwort</i>
<i>stammen</i>	<i>Anderer Mensch</i>	<i>Team</i>	<i>Danke Otto</i>
<i>halten</i>	<i>Dunkle Hautfarbe</i>	<i>The</i>	<i>Klares Statement</i>
<i>verweigern</i>	<i>Offener Brief</i>	<i>My</i>	<i>Klare Kante</i>
<i>schlimm</i>	<i>Großes Problem</i>	<i>Kante</i>	<i>Neuer Kunde</i>
<i>kotzen</i>	<i>Normaler Mensch</i>	<i>Reagieren</i>	<i>Lieber Mensch</i>
<i>null</i>	<i>Braune Gesinnung</i>	<i>Marketing</i>	<i>Liebes Team</i>

TABLE 1.4 Fifty keywords for the Otto-Twitter corpus in comparison to the web corpus déteinte.

Lemma in German	English	Frequency	Keyness-Score	Type	Lemma in German	English	Frequency	Keyness-Score	Type
1 <i>anwidern</i>	disgust	82	280,2	1	26 <i>dünnschiss</i>	the shits	7	53,1	1
2 <i>rassist</i>	racist	87	253,0	3	27 <i>abschaum</i>	scum	10	52,9	1
3 <i>hautfarbe</i>	skin color	97	200,6	3	28 <i>scheiße</i>	shit	34	52,9	1
4 <i>dunkelhäutig</i>	dark-skinned	38	173,6	3	29 <i>afd</i>	afd	126	48,7	4
5 <i>blocken</i>	block	109	168,8	2	30 <i>biodeutsche</i>	bio-german	6	48,6	3
6 <i>widerlich</i>	disgusting	64	162,1	1	31 <i>widerwärtig</i>	disgusting	10	48,2	1
7 <i>fascio</i>	fascist	22	157,2	4	32 N-Wort	n-word	13	45,3	3
8 <i>ekelhaft</i>	despicable	45	147,1	1	33 <i>arier</i>	arian	7	42,7	4
9 <i>blockliste</i>	blocklist	18	146,6	2	34 <i>liken</i>	like	6	42,7	2
10 <i>rassistin</i>	racist	17	136,6	3	35 <i>muhen</i>	moo	7	41,8	Other
11 <i>dito</i>	ditto	33	136,4	6	36 <i>kotzen</i>	vomit	18	41,0	1
12 <i>tweets</i>	tweets	40	114,5	2	37 <i>likes</i>	likes	13	40,1	2
13 <i>glotzen</i>	gawk	20	105,3	Other	38 <i>poc</i>	poc	6	39,8	3
14 <i>rassistisch</i>	racist	122	101,8	3	39 <i>illegale</i>	illegal	7	39,7	3
15 <i>nazi</i>	nazi	169	100,4	4	40 <i>farbige</i>	colored	7	39,6	3
16 <i>altpartei</i>	old party	14	79,1	5	41 <i>volksverhetzung</i>	incitement of people	9	39,4	4
17 <i>fascokuschler</i>	fascist cuddler	9	75,7	4	42 <i>negatiubeispiel</i>	negative example	6	38,9	6
18 <i>rassismus</i>	racism	114	72,5	3	43 <i>hirn</i>	brain	34	38,5	Other
19 <i>nazis</i>	nazis	10	70,6	4	44 <i>finden</i>	find	104	38,5	6
20 <i>scheiße</i>	shit	65	69,1	1	45 <i>migrationspakt</i>	migration pact	6	38,4	5
21 <i>afrikaner</i>	africans	25	65,4	3	46 <i>bravo</i>	bravo	18	38,3	6
22 <i>blockieren</i>	block	102	59,4	2	47 <i>gedankengut</i>	thinking	16	38,0	6
23 <i>Aluhut</i>	Aluminium hat	8	59,3	4	48 <i>chapeau</i>	chapeau	8	38,0	6
24 <i>arisch</i>	aryan	12	58,5	4	49 <i>hetzen</i>	stir up hatred	21	37,6	4
25 <i>fascist</i>	fascist	22	56,0	4	50 <i>umvolkung</i>	umvolkung	5	37,5	5

TABLE 1.5 Key Discourse Types in Table 1.4.

-
1. Outrage and devaluation
 2. Social media practices / counterspeech practices
 3. Addressing Racism and Racism-Relevant Membership Categorization Devices
 4. Anti-discriminatory discourse: fascism, right-wing populism, conspiracy
 5. Terms from right-wing discourse spectrum
 6. Opinion and evaluation
-

patterns at the lexical level, visible in the recurrent labeling practices for Paula and her followers in the form of membership categorization devices such as fascist, racist, or Aryan. This is an anti-discriminatory interactional practice, which Whitehead (2009) described as “categorizing the categorizer”. The mirroring of indignation and emotional-affective language use in the initial tweet, visible in *disgust*, *vomit*, *shit*, or *despicable* but also approval practices (*ditto*), evaluation practices (*negative example*, *find*), and praise for Otto’s reply (*chapeau*, *bravo*) provide an initial orientation and clues about the thematic and interactional course of the comment column.

Shifting from a corpus-linguistic to a microanalytical perspective, we examine Otto’s response comment, which serves as an excellent illustration of the various linguistic means employed to establish opposition to the trigger event. The foundational principle of the mirroring practice laid out below hinges on resumption and structure adoption. Construction borrowing techniques have been primarily described in text linguistic studies as isotopy chains or semantic contiguities for coherence creation, achieved through the repetition of identical or adjacent lexemes. The concept of “structure adoption” goes a step further, proposing that two conversational steps are linked by adopting the syntactic structure of the preceding step. This approach allows us to distinguish between partial and total adoption practices.

Although it is generally claimed that this procedure serves as a means of cooperativity and the display of agreement/concordance (see Brinker and Sager 2010), the contradictory quality of such takeovers is thereby omitted. For example, Androutsopoulos (2019) used editorial moderation of comment columns to show how construction borrowing can be a tried-and-true means of posing oneself and others in the debate in antagonistically shaped debates. In cognitive linguistic terms, the construction borrowing and alignment of lexical or syntactic structures is also called interactive alignment, assuming that through the same representations on the linguistic surface (at different linguistic levels), interaction participants thus “arrive at similar conceptual ideas through a shared vocabulary and through shared syntactic structures” (Lotze 2014, 279).

I will now show how the procedure of taking over constructions is used for exactly the opposite purpose, namely, that by rearranging and substituting certain elements of the previous utterances, contradictory ideas are to be expressed.

Paula's initial tweet (1) (see Table 1.1) is the framework that Otto employs and where certain elements, which are to be communicatively countered, are substituted. On the surface, the tweet sends signals of co-indignation about the practice and suggests agreement via the twice repeated "also". By substituting and erasing certain elements, the contradiction emerges in contrast to Paula's tweet, and, at the same time, explicit counter-assertions are made with it (using almost the identical syntactic and lexical means). "Through repetition, the proposition in question gains meta-pragmatic attention and is, in a sense, exposed" (Androutsopoulos 2019, 280). Often, only a single lexical variation with a change of semantic paradigm (e.g., from Africa to Tweet) is sufficient (see Androutsopoulos 2019). With this strategic approach, Otto skillfully avoids explicitly commenting on the implicit accusations in Paula's tweet while simultaneously refraining from alluding to the categories Paula implicates. Initial signs of emotional alignment are discernible, emphasizing shared disgust – and thus taking over the expressive speech act. Instead, Otto focuses on Paula's tweet, replacing the racism-relevant category "Africa" from "Afrika glotzt mich an" with "uns glotzt erneut so ein absurder Tweet an". In this way, Otto avoids Paula's provocation without having to explicitly refute her claims about Africa.

The omission or deletion of an equivalent for "society we should become" enables Otto to reject the accusation of political interests or of a targeted mass media legitimization of population exchange through commercial advertising. While "This is not our society" does not undergo any formal variation, the new deictic reference structure creates an alternative assertion: persons with racist attitudes are not part of the society Otto considers desirable. The demarcation "without us" merely modifies the pro-form to "us". In contrast to the last part of the initial tweet, the rainbow emojis could be interpreted as a positive evaluation of diversity.

In essence, Otto's tweet now serves as a creative and interpretive resource, providing the inferential apparatus for the practical deconstruction of racist formulations. Moreover, such phrases have been labeled and recognized as potentially racist by members of the social media speech community.

Example (3) in Table 1.6 illustrates a tweet by User_ComX, which is one of many that uses extensive construction borrowing and thus goes into syntagmatic

TABLE 1.6 Transcript of example three of a tweet by User_ComX.

(3)	2↔ @paulapaulinchen @otto_de Da will man einen schönen Tag haben, geht bei Twitter rein und dann klotzt mich ein Mitglied der AfD an. ICH WILL DAS NICHT MEHR SEHEN!!! AfD ist nicht unsere Gesellschaft, dass ist die Gesellschaft die wir nicht wollen. OHNE MICH! @paulapaulinchen find ich scheiße	User_ComX (TW-OTTO-0100) 🕒14.02.2020 13:45:15 🗨️2/♥️105/👍4
	@paulapaulinchen @otto_de You want to have a nice day, login on Twitter, and an AfD member stares at you! I DON'T WANT TO HAVE TO SEE THIS ANYMORE! AfD is not our society, this is the society we're not meant to become. WITHOUT ME! @paulapaulinchen I think you suck.	

and paradigmatic relation in opposition to Paula's initial tweet. This pattern is extremely productive in the present data and occurs in multiple variations.

Briefly summarized: This section analyzed the case study Otto vs. Paula, in which Paula expresses her dislike of the portrayal of black models on the website of the Otto company. In doing so, she refers to her racist views and attitudes toward social change, which she opposes. Paula's tweet serves as a template ("template") for a series of responses ("counterspeech") that take up content and linguistic parameters of the original tweet to express dissent. Otto's response uses the initial tweet as a template for his counterspeech practice. In doing so, Otto uses the process of "construction takeover" by rearranging parts of Paula's tweet and substituting certain elements to express an opposing stance. The analysis also shows that the discourse is mainly dominated by counterspeech and identifies several recurring themes and discourse strands. Finally, a microanalytical perspective on Otto's reply tweet is presented, showing the different linguistic means by which the tweet establishes an opposition to the initial tweet.

Offer Alternative Positions and Change Positions

As a second example, I will now analyze a multistep hate-counterspeech dynamic, a so-called counter-conversation.

In example (4), as can be seen in Table 1.7, User_Mi continues Otto's counterspeech strategy by challenging Paula's presuppositions. He uses a targeted question to question Paula's assumptions about the depicted model: "I'm much more interested in how you know this beautiful young woman is African?" He questions Paula's ability to determine a person's nationality based on their skin color, introducing the idea that black people can also be German. This is a crucial interactional reversal: User_Mi redefines what it means to be "German", thereby questioning Paula's presupposition. User_Hei, a follower of Paula, enters the interaction to defend Paula's statement, interpreting it as a critique of "multi culti" propaganda, a term commonly used in right-wing conservative and far-right circles. However, this attempt to position himself on Paula's side is challenged by User_CA. CA uses negations to reject this interpretation and make a counter-positioning clear. Here we see the use of linguistic tools like negations and the ironic use of member categorization devices to counteract User_Hei's views. By ironically picking up the member categorization device "white, blonde, blue-eyed people", CA attempts to expose the underlying assumptions about appropriate advertising models and homogeneity notions in Germany. This shows the indexical nature of communicative practices, where linguistic means refer to a situation in which right-wing populist and racist notions are processed. This makes the practice of counterspeech visible to all participants and researchers.

User_Hei tries to control the ongoing discourse by presenting a radical vision of the future that he portrays as inevitable: a society that enforces multiculturalism.

TABLE 1.7 Transcript of counter-conversation “Germans are not only white, blond, blue-eyed people”.

(4)	2↳	<p>@Paula @otto_de Mich würde viel eher Interessieren, woher du weißt, dass diese hübsche junge Frau Afrikanerin ist? Ich weiß, dass übersteigt deinen Horizont aber es gibt farbige/Mischlinge, die hier in Deutschland geboren worden sind nur deutsch sprechen. Sie sind DEUTSCHE.</p> <p>@Paula @otto_de I would be much more interested in how you know that this pretty young woman is African? I know this is beyond your horizon but there are coloured/half-castes[mulattos] who were born here in Germany and only speak German. They are GERMAN.</p>	<p>User_Mi (TW-OTTO-0102) 🕒14.02.20 10:57:47 👍0/👎35/👤5</p>
(5)	↳	<p>@User_Mi @Paula @otto_de Schön, wie man es aus dem Kontext reißen kann. Es geht im Grundsatz darum, dass selbst in der Werbung Multikulti propagiert wird. Es hat den Anschein, als müsse es nur lange genug gezeigt werden. Es kann nur nicht so funktionieren.</p> <p>@User_Mi @Paula @otto_de Nice how you can take it out of context. The basic point is that even in advertising multiculti is propagated. It feels like it just needs to be shown long enough. It just can't work that way.</p>	<p>User_Hei (TW-OTTO-0203) 🕒14.02.20 12:11:00 👍0/👎0/👤1</p>
(6)	↳	<p>@User_hei @User_Mi @Paula @otto_de Ne, es ist viel einfacher: Deutsche sind schon längst nicht nur weiße, blonde, blauäugige Menschen - waren es vermutlich noch nie. Ergo zeigt die Werbung eben auch nicht nur weiße, blonde, blauäugige Menschen. Da muss man gar nichts "propagieren". So ist "das Volk"</p> <p>@User_hei @User_Mi @Paula @otto_de No, it's much simpler: Germans are not only white, blond, blue-eyed people - they probably never were. Ergo, advertising does not only show white, blond, blue-eyed people. There is no need to "propagate" anything. That's how "the people" are</p>	<p>User_CA (TW-OTTO-0274) 🕒14.02.20 13:03:19 👍0/👎3/👤1</p>
(7)	↳	<p>@User_CA @User_NY @Paula @otto_de So soll es werden, falls Du es noch nicht bemerkt hast. Ich halte das trotzdem für genauso schwachsinnig, wie die Nürnberger Rassengesetze. Aufgezwungen wird immer Probleme geben.</p> <p>@User_CA @User_NY @Paula @otto_de That's how it should become, in case you haven't noticed. I still think it's as moronic as the Nuremberg racial laws. Imposed will always cause problems.</p>	<p>User_Hei (TW-OTTO-0280) 🕒14.02.20 13:06:53 👍0/👎0/👤2</p>
(8)	↳	<p>@User_Heiz @User_NY @Paula @otto_de Nee, nochmal zum Mitschreiben: Es IST bereits so. Und zwar schon lange. Hier leben Menschen mit verschiedenen Hautfarben. Und das sind tatsächlich richtige, echte, gebürtige Deutsche, die mit dem Sandmännchen, Günther Jauch, dem FC Bayern und Goethes Faust aufwachsen.</p>	<p>User_CA (TW-OTTO-0295) 🕒14.02.20 13:18:59 👍0/👎1/👤1</p>

- @User_Heiz @User_NY @Paula @otto_de Nah, again for the record: It IS already like this. And it has been for a long time. People with different skin colours live here. And these are actually real, genuine, native Germans who grew up with Sandmännchen, Günther Jauch, FC Bayern and Goethe's Faust.
- (9) ↳ @User_CA @User_NY @Paula @otto_de Es geht nicht um den Teil, der ist! Es geht um den, der aufgezwungen werden soll. Das ist Multikulti um jeden Preis und das wird im Krieg enden. Die Werbung soll nur dahingehend indoktrinieren.
- @User_CA @User_NY @Paula @otto_de It's not about the part that is! It's about the one that is to be imposed. It's multiculti at any cost and it will end in war. The advertising is only meant to indoctrinate.
- (10) ↳ @User_Heiz @User_CA @User_NY @Paula @otto_de Wenn du willst, schick ich dir ne Bauanleitung für einen Aluhut.
- @User_Heiz @User_CA @User_NY @Paula @otto_de If you want, I'll send you instructions on how to make an tinfoil hat.
- (11) ↗ @User_Hei @User_CA @User_NY @Paula @otto_de "Propagieren von Multikulti" = Nürnberger Rassengesetze. Völlig logisch und kein bisschen geschichtsrevisionistisch. Wo du schon Serdar als "Unterstützer" deines Schwachsinnnes heranziehst: Wenn er deinen Feed lesen würde, müsste er vermutlich kotzen.
- @User_Hei @User_CA @User_NY @Paula @otto_de "Propagating multiculturalism" = Nuremberg racial laws. Completely logical and not a bit historical revisionist. Since you're already using Serdar as a "supporter" of your bullshit: If he would read your feed, he would probably puke.
- User_Hei
(TW-OTTO-0322)
⌚14.02.20 13:33:38
👍0/👎0/💬2
- User_Li
(TW-OTTO-0334)
⌚14.02.20
13:40:30 👍0/👎1/💬1
- User_GS
(TW-OTTO-2012)
⌚15.02.20 09:16:53
👍0/👎0/💬0

This notion is depicted as absurd, compared to the Nuremberg Race Laws. Here we see a typical example of the so-called slippery slope argument, which presents an extreme, undesirable outcome as an inevitable consequence of a particular policy. User_CA responds again with a counter-positioning. Instead of focusing on the alleged forced future, he emphasizes the current reality: Germany is already multicultural. The application of the present tense and the enumeration of shared cultural elements (Sandmännchen, Günther Jauch, FC Bayern, and Goethe's Faust) serves to highlight the shared national heritage and emphasize that people of different skin colors participate in it. User_Hei insists on his position that multiculturalism is being forced and will end in war. In this statement, we find a repeated use of the slippery slope argument, but this time with an intensified emotional and alarming statement predicting the end in war. The introduction of

User_Li adds a humorous and ironic element to the discussion by offering User_Hei a “building instruction for a tin foil hat”. The “tin foil hat” is a well-known symbol for conspiracy theories and is often used to ridicule belief in such theories and to label someone as a conspiracy theorist. This undermines User_Hei’s arguments by portraying them as paranoid conspiracy theories.

In the end, User_GS enters the dialogue and sharply criticizes User_Heiz for his equivalence between the promotion of multiculturalism and the Nuremberg Race Laws, which he portrays as revisionist and irrational (see example (11) in Table 1.7). He also questions User_Heiz’s credibility by suggesting that a supporter cited by User_Heiz, Serdar, would be disgusted if he read User_Heiz’s opinions.

In the presented Twitter conversation, there are clear indications that counterspeech works as a collective and collaborative practice. Different counterspeakers use various strategies and approaches to confront the same idea. While User_GS uses direct confrontation and discrediting, User_CA and User_Mi use astute questions and hypothetical scenarios to steer the discussion in new directions and highlight the complexity of identity and nationality.

Moreover, the counterspeakers complement each other in the way they present their arguments. User_Heiz’s extreme case formulations on the Nuremberg Race Laws are picked up and “overheard” by User_CA to deepen the original discussion and further question User_Heiz’s original views. User_GS joins the debate later, about a day later, after he has read the entire thread and realizes that no one has yet responded to the comparison to the Nuremberg Race Laws, relativizing National Socialism, and makes up for it. This collective participation enables the discussion to be brought to a broader and deeper level than a single speaker could.

Finally, the counterspeakers manage to collectively create a dominant voice of resistance against User_Heiz’s arguments. Through their different approaches and their joint participation in the discussion, they present a strong collective front that challenges the original views and invites other voices to join them.

This dialogue illustrates how counterspeech operates not only at the level of the *what* but also at the level of the *how* – how linguistic means and interactive, sometimes collaborative procedures emerge and come into play in practice. This analysis opens up a more complex and richer conception of counterspeech that goes beyond simplified notions and enriches existing research.

Conclusion

The definition and the theory of counterspeech I proposed involve both limitations and extensions regarding existing approaches. Not every comment that is directed against an antecedent is considered counterspeech. Instead, it requires the management of invective situations to assert anti-discriminatory discourse positions and/or ensure the protection of those insulted, based on normative and moral values which are inscribed into certain communicative practices (as ideology; cf. Hanks 1996). However, it also enables the description of invective counterspeech and is not solely confined to civil forms of counterspeech such as fact-

checking and counter-arguing. What matters is the application and effect of the invective counterspeech and how it is discussed in the follow-up communication.

Examining counterspeech as an interaction needs an understanding of counter-conversations as a cooperative endeavor involving at least two participants. In discussions about counterspeech and hate speech, the task involves responding to perceived racist remarks and minimizing their impact. This is an extremely complex task requiring a variety of linguistic tools and communicative strategies that can be employed depending on their situational and media context, significantly influenced by the previous interaction history.

In the present case study *Otto vs. Paula*, the initiating tweet, as well as Paula's profile and activity history, constituted the most crucial resources for contradiction not only because thematic content aspects were revisited but also because the takeover procedure at various linguistic levels impacted the subsequent tweets. The process of taking over and targeted substitution, rearrangement of individual parts, or complete obliteration was identified as a frequently and variably utilized procedure of contradiction, creating new reference structures that can be described as opposing or alternative claims in relation to the initiating tweet.

For the future, it seems desirable to analyze further case examples and investigate the resource bundles of counterspeech practice to differentiate the linguistic-communicative inventory for specific contexts and counterspeech strategies.

Notes

- 1 See Caponetto and Cepollaro (chapter 2) as well as Bahador (chapter 4) in this volume for more detailed discussions of counterspeech strategies and tactics.
- 2 See Costa, Mendes da Silva, and Tavares (chapter 6) as well as Tomalin, Roy, and Weisz (chapter 7) in this volume for more information on the CONAN data set.
- 3 See also Saltman and Munir (chapter 8) in this volume.
- 4 For example, searching for the term *Gegenrede* on German YouTube leads to an "alternative" talk show by the Alternative for Germany, a right-wing party in Germany.

References

- Álvarez-Benjumea, Amalia, and Fabian Winter. 2018. "Normative Change and Culture of Hate: An Experiment in Online Environments." *European Sociological Review* 34 (3): 223–237. doi:10.1093/esr/jcy005
- Androutsopoulos, Jannis. 2019. "Wiederaufnahmen im Nutzerdialog: Eine medienlinguistische Fallstudie zur Bewältigung vernetzter Interaktion im digitalen Journalismus." In *Interaktion und Medien: Interaktionsanalytische Zugänge zu medienvermittelter Kommunikation*, edited by Konstanze Marx and Axel Schmidt, 1st ed., 257–285. OraLingua17. Heidelberg: Universitätsverlag Winter.
- Austin, John Langshaw. 1962. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*. London: Oxford University Press.
- Bahador, Babak. 2021. "Countering Hate Speech." In *The Routledge Companion to Media Disinformation and Populism*, edited by Howard Tumber and Silvio Waisbord, 507–518. New York: Routledge.

- Baider, Fabienne, and Monika Kopytowska. 2018. "Narrating Hostility, Challenging Hostile Narratives." *Lodz Papers in Pragmatics* 14 (1): 1–24. doi:10.1515/lpp-2018-0001
- Bartlett, Jamie, and Alex Krasodomski-Jones. 2015. *Counter-Speech: Examining Content That Challenges Extremism Online*. <https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf>. Accessed 8 September 2018.
- Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. *Counterspeech on Twitter: A Field Study*. <https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/>. Accessed 1 July 2020.
- Bick, Eckhard, Klaus Geyer, and Andrea Kleene. 2021. „Die Äch so Friedlichen Muslime“: Eine Korpusbasierte Untersuchung Von Formulierungsmustern Fremdenfeindlicher Aussagen in Sozialen Medien." In *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen: Theoretische Und Empirische Annäherungen an Ein Interdisziplinäres Phänomen*, edited by Sebastian Wachs, Barbara Koch-Priewe, and Andreas Zick, 81–103. Wiesbaden: Springer Fachmedien Wiesbaden.
- Brinker, Klaus, and Sven F Sager. 2010. *Linguistische Gesprächsanalyse: Eine Einführung*. 5th ed. Grundlagen der Germanistik30. Berlin: Erich Schmidt.
- Butler, Judith. 2006. *Haß spricht: Zur Politik des Performativen*. Edition Suhrkamp 2414. Frankfurt am Main: Suhrkamp.
- Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. "CONAN – COUNTER NARRATIVES THROUGH NICHE-SOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 2819–2829. Cedarville, OH: Association for Computational Linguistics. doi:10.18653/v1/P19-1271
- Deppermann, Arnulf, Helmuth Feilke, and Angelika Linke. 2016. "Sprachliche und kommunikative Praktiken: Eine Annäherung aus linguistischer Sicht." In *Sprachliche und kommunikative Praktiken*, edited by Arnulf Deppermann, Helmuth Linke, and Angelika Feilke, 1–24. Berlin: De Gruyter.
- Downs, Donald A. 1985. "Skokie Revisited: Hate Group Speech and the First Amendment." *Notre Dame Law Review* 60 (4): 629–685. <https://scholarship.law.nd.edu/ndlr/vol60/iss4/1>. Accessed 5 September 2023.
- Durrheim, Kevin, Ross Greener, and Kevin A Whitehead. 2015. "Race Trouble: Attending to Race and Racism in Online Interaction." *The British Journal of Social Psychology* 54 (1): 84–99. <https://doi.org/10.1111/bjso.12070>
- Digital Dictionary of the German Language*. 2023. "Gegenrede, die." <https://www.dwds.de/wb/Gegenrede>. Accessed 5 September 2023.
- Ellerbrock, Dagmar, Lars Koch, Sabine Müller-Mall, Marina Münkler, Joachim Scharloth, Dominik Schrage, and Gerd Schwerhoff. 2017a. *Invektivität – Perspectives for a New Research Program in Cultural Studies and the Social Sciences*. https://tu-dresden.de/gsw/sfb1285/ressourcen/dateien/Invektivitat_-English-_Anm_GP_final.pdf/at_download/file
- Ellerbrock, Dagmar, Lars Koch, Sabine Müller-Mall, Marina Münkler, Joachim Scharloth, Dominik Schrage, and Gerd Schwerhoff. 2017b. "Invektivität – Perspektiven eines neuen Forschungsprogramms in den Kultur- und Sozialwissenschaften." *Kulturwissenschaftliche Zeitschrift* 2 (1): 2–24. doi:10.2478/kwg-2017-0001
- Fedtko, Cornelia, and Gregor Wiedemann. 2020. "Hass- und Gegenrede in der Kommentierung massenmedialer Berichterstattung." In *Soziale Medien: Interdisziplinäre Perspektiven Auf Social Media*, edited by Peter Klimczak, Christer Petersen, and Samuel Schilling, 91–120. ars digitalis. Wiesbaden: Springer Fachmedien Wiesbaden.

- Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martínez. 2015. *Countering Online Hate Speech*. <https://unesdoc.unesco.org/ark:/48223/pf0000233231>. Accessed 19 February 2022.
- Gensing, Patrick. 2010. "Widerständige Struktur: Kommentar Neonazis bei Facebook." *taz*, 14 May 2010. <https://taz.de/Kommentar-Neonazis-bei-Facebook/15142677/>.
- Geyer, Klaus, Eckhard Bick, and Andrea Kleene. 2022. "I Am No Racist but" In *The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse*, edited by Natalia Knoblock, 41–261. Cambridge: Cambridge University Press.
- Giles, David. 2021. "Context, History, and Twitter Data: Some Methodological Reflections." In *Analysing Digital Interaction*, edited by Joanne Meredith, David Giles, and Wyke Stommel, 41–63. Palgrave Studies in Discursive Psychology. Cham: Palgrave Macmillan.
- Hanks, William F. 1996. *Language & Communicative Practices*. Critical Essays in Anthropology. Boulder, CO: Westview Press.
- Hoffmann, Ludger. 2020. "Zur Sprache Des Rassismus." *Sprachreport* 36 (1): 40–47.
- Holzleithner, Elisabeth. 2000. "Grenzziehungen: Pornographie, Recht und Moral." Diss., Universität Wien.
- Housley, William, Helena Webb, Adam Edwards, Rob Procter, and Marina Jirotko. 2017. "Membership Categorisation and Antagonistic Twitter Formulations." *Discourse & Communication* 11 (6): 567–590. doi:10.1177/1750481317726932
- Housley, William, Helena Webb, Meredydd Williams, Rob Procter, Adam Edwards, Marina Jirotko, Pete Burnap, Bernd Carsten Stahl, Omer Rana, and Matthew Williams. 2018. "Interaction and Transformation on Social Media: The Case of Twitter Campaigns." *Social Media + Society* 4 (1). doi:10.1177/2056305117750721
- Hufer, Klaus-Peter. 2016. *Argumente Am Stammtisch: Erfolgreich Gegen Parolen, Palaver Und Populismus*. 7th ed. Schwalbach/Ts: Wochenschau.
- Imo, Wolfgang. 2016. "Im Zweifel für den Zweifel: Praktiken des Zweifelns." *Sprachliche und kommunikative Praktiken*, edited by Arnulf Deppermann, Helmuth Linke, and Angelika Feilke, 153–176. Berlin: De Gruyter.
- jugendschutz.net. 2013. *Rechtsextremismus Online Beobachten und Nachhaltig Bekämpfen: Bericht Über Recherchen und Maßnahmen im Jahr 2012*. https://www.hass-im-netz.info/fileadmin/public/main_domain/Dokumente/Rechtsextremismus/Rechtsextremismus_online_2012.pdf. Accessed 26 June 2020.
- Kalven, Harry. 1960. "The Metaphysics of the Law of Obscenity." *The Supreme Court Review* 1960: 1–45. <http://www.jstor.org/stable/3108685>
- Klinker, Fabian, Joachim Scharloth, and Joanna Szczek, eds. 2018. *Sprachliche Gewalt: Formen und Effekte von Pejorierung, verbaler Aggression und Hassrede*. Abhandlungen zur Sprachwissenschaft. Stuttgart: J.B. Metzler. doi:10.1007/978-3-476-04543-0
- Kuck, Kristin. 2019. "Argumentieren Gegen Rassismus – Inszenierung Und Strategie." *Sprachreport* 35 (1): 32–43.
- Langton, Rae. 2018. "The Authority of Hate Speech." In *Oxford Studies in Philosophy of Law*, edited by John Gardner, Leslie Green, and Brian Leiter, vol. 3, 123–152. Oxford: Oxford University Press. doi:10.1093/oso/9780198828174.003.0004
- Lanwer, Jens, and Georgios Coussios. 2018. "7. Kommunikative Praxis, soziale Gruppe und sprachliche Konventionen." In *Handbuch Sprache in Gruppen*, edited by Eva Neuland and Peter Schlobinski, 126–148. Handbücher Sprachwissen Band 9. Berlin: De Gruyter.
- Linell, Per. 2012. "Zum Begriff Des Kommunikativen Projekts." In *Sozialität in Slow Motion*, edited by Ruth Ayaß and Christian Meyer, 71–79. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Lotze, Netaya. 2014. "Interaktives Alignment im Dialog." In *Sprachen? Vielfalt! Sprache und Kommunikation in der Gesellschaft und den Medien. Eine Online-Festschrift zum Jubiläum von Peter Schlobinski*, edited by Alexa Mathias, Jens Runkehl, and Torsten Siever, 275–288. Hannover: Leibniz Universität.
- Marx, Konstanze. 2020. "Warum automatische Verfahren bei der Detektion von Hate Speech nur die halbe Miete sind." In *Cyberkriminologie: Kriminologie für das digitale Zeitalter*, edited by Thomas-Gabriel Rüdiger and Petra S Bayerl, 707–725. Wiesbaden: Springer Fachmedien Wiesbaden.
- Marx, Konstanze, and Sebastian Zollner. 2020. "Counter Speech in sozialen Medien: Strategien Digitaler Zivilcourage erlernen und anwenden." *Deutsch* 5–10 (63): 24–28.
- Matsuda, Mari J. 1989. "Public Response to Racist Speech: Considering the Victim's Story." *Michigan Law Review* 87 (8): 2320–2381. doi:10.2307/1289306
- Meibauer, Jörg, ed. 2013. *Hassrede/Hate Speech: Interdisziplinäre Beiträge Zu Einer Aktuellen Diskussion*. Giessen: Giessener Elektronische Bibliothek.
- Meibauer, Jörg. 2022. *Sprache und Hassrede*. Kurze Einführungen in die germanistische Linguistik Band 29. Heidelberg: Universitätsverlag Winter.
- Meier-Vieracker, Simon. 2020. "Selbstlegitimationen von Hass auf rechten Internetseiten." In *Diskurs – Ethisch*, edited by Heidrun Kämper and Ingo H Warnke, 139–155. Bremen: Hempen.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, and Peter Norvig. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science (New York, N.Y.)* 331 (6014): 176–182.
- Mildorf, Jarmila, and Bronwen Thomas. 2017. "Introduction." In *Dialogue across Media*. Dialogue Studies (DS), edited by Jarmila Mildorf and Bronwen Thomas, vol. 28, 1–16. Amsterdam: John Benjamins.
- Miškolci, Jozef, Lucia Kováčová, and Edita Rigová. 2020. "Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia." *Social Science Computer Review* 38 (2): 128–146. <https://doi.org/10.1177/0894439318791786>
- Oxford English Dictionary. 2023a. "Counter-, prefix." https://www.oed.com/dictionary/counter_prefix?tab=meaning_and_use#8069795. Accessed 5 September 2023.
- Oxford English Dictionary. 2023b. "Counter-speech, noun." https://www.oed.com/dictionary/counter-speech_n?tab=meaning_and_use#8098996. Accessed 5 September 2023.
- Rieger, Diana, Josephine B Schmitt, and Lena Frischlich. 2018. "Hate and Counter-Voices in the Internet: Introduction to the Special Issue." *SC|M* 7 (4): 459–472. doi:10.5771/2192-4007-2018-4-459
- Selting, Margret. 2016. "Praktiken des Sprechens und Interagierens im Gespräch aus der Sicht von Konversationsanalyse und Interaktionaler Linguistik." In *Sprachliche und kommunikative Praktiken*, edited by Arnulf Deppermann, Helmuth Linke, and Angelika Feilke, 27–56. Berlin: De Gruyter.
- SFB Invektivität. 2021. "*Forschungsprogramm & Glossar Zentraler Begriffe*." <https://tu-dresden.de/gsw/sfb1285/forschung/forschungsprogramm>. Accessed 11 April 2022.
- Strossen, Nadine. 1997. "Why the American Civil Liberties Union Opposes Campus Hate Speech Codes." *Academic Questions* 10 (3): 33–40.
- Strossen, Nadine. 2018. *Hate: Why We Should Resist It with Free Speech, Not Censorship*. Inalienable Rights Series. New York: Oxford University Press.
- Sunstein, Cass R. 1986. "Pornography and the First Amendment." *Duke Law Journal* 4: 589–627. <https://heinonline.org/HOL/P?h=hein.journals/duk1986&ci=615>. Accessed 5 September 2023.

- Thürer, Daniel. 2007. *Rassismus Als Gift – Reaktion Durch Das Recht*. <https://www.nzz.ch/articleEWTMD-ld.397482>. Accessed 24 April 2023.
- Warnke, Ingo H. 2021. *Widerrede als Fürsprache? Überlegungen zur Doppelgesichtigkeit von Empörung*. <https://www.wiko-greifswald.de/mediathek/beitrag/n/widerrede-als-fuersprache-ueberlegungen-zur-doppelgesichtigkeit-von-empörung-87847/>. Accessed 19 February 2022.
- Whitehead, Kevin A. 2009. ““Categorizing the Categorizer”: The Management of Racial Common Sense in Interaction.” *Social Psychology Quarterly* 72 (4): 325–342. doi:10.1177/019027250907200406
- Wright, Lucas, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. “Vectors for Counterspeech on Twitter.” In *Proceedings of the First Workshop on Abusive Language Online*, edited by Zeerak Waseem, Wendy HK Chung, Dirk Hovy, and Joel Tetreault, 57–62. Stroudsburg, PA: Association for Computational Linguistics.
- Zollner, Sebastian. 2022. “Counter Speech als sprachlich-kommunikative Praktik in digitalen, invektiven Konstellationen: Ein Thema für Linguistik, Medienpädagogik und politische Bildung.” *merz. medien + erziehung* 66 (2): 35–45.

2

THE PHILOSOPHY OF COUNTER LANGUAGE

Laura Caponetto and Bianca Cepollaro

Introduction

Speech can be “toxic”, as philosopher Lynne Tirrell, among others, has recently emphasized (Tirrell 2017, 2018). Tirrell provided a broad characterization of *toxic speech* as speech that “diminishes democratic participation, undermines civil liberties, compromises the rule of law, and damages human dignity” (Tirrell 2018, 120). We will here use the term to refer to speech that spreads prejudicial stereotypes and/or endorses discriminatory practices (such as sexism, racism, homophobia, transphobia, ableism, etc.). The epidemiological metaphor aptly suggests that toxic speech can operate implicitly, rather than explicitly. Just as toxins silently stockpile, eroding the body little by little, so, too, toxic utterances can cumulatively disrupt the social fabric by propagating discriminatory attitudes in surreptitious ways. Toxic speech, so construed, includes both blatantly hateful utterances and subtler forms of discriminatory discourse and thus forms a broader category than *hate speech*.¹

Scholars have pointed out that toxic speech has the potential to shape our epistemic and normative landscapes, by changing what we believe and accept, as well as what is permissible in a given context (see, e.g., McGowan 2009, 2019; Langton 2012, 2018; Caponetto and Cepollaro 2021). The claim is supported by psychological evidence showing that derogatory language, and speech toxicity more broadly, reduces the well-being of and increases suicide rates among targeted individuals (Swim et al. 2001; Mullen and Smyth 2004; Leader, Mullen, and Rice 2009), while prompting implicit negative evaluations and dehumanizing attitudes toward them (Carnaghi and Maass 2007; Fasoli, Maas, and Carnaghi 2015; Fasoli et al. 2016; Soral, Bilewicz, and Winiewski 2018). Against this backdrop, the question of how to resist becomes particularly pressing. If toxic speech can have harmful effects on individuals and the social

world, then devising strategies to counter it is of paramount importance. This is where the debate on counterspeech starts out.

“Counterspeech” is a term of art introduced in legal theory to pick out a family of measures that are alternative to censorship. As Justice Brandeis famously put it in *Whitney v. California*, the remedy to toxic speech (or “evil” speech, as he called it) should be “more speech, not enforced silence” (Brandeis 1927). But what forms should “more speech” (or counterspeech) take? Philosophy of language has developed tools and drawn distinctions that can shed light on the forms that toxic speech can take and the most suitable ways to tackle them. As we shall see, engineering counterspeech to suit the communicative features of the toxic utterance it responds to may increase its chances to hit the mark.

This chapter provides an opinionated survey of a number of counterspeech strategies that have been variously discussed in contemporary philosophy of language.² We will point out that certain strategies are particularly apt to counter toxic contents that are conveyed *implicitly*, whereas others have their best shot with toxic contents that are *explicitly* stated. We will also suggest that the appropriateness and expected outcome of a given strategy importantly vary with the context and that the counterspeaker’s role (e.g., their belonging to the targeted group or not; their speaking as a private citizen or a government representative) is one of the major contextual variables.³ Overall, our goal is to uncover how tools from philosophy of language can illuminate the workings of toxic speech and help us devise strategies to counter it. In this pursuit, we will sketch the foundations of a philosophy of counter-language.

Counterspeech Strategies

Denying

The most intuitive strategy to counter toxic speech through more speech is to reject it as false, possibly by providing reasons or evidence against it. To see the strategy at work, suppose that John, Paul, and Arthur are chitchatting, when they see Sarah, an acquaintance of theirs, in a large SUV across the street. John says,

1 Wow! That’s huge. No way she can park it. Women just can’t drive.

John’s utterance is clearly toxic: it contributes to spreading, and openly endorses, the idea that “women can’t drive” – a sexist stereotype that is part and parcel of a system of representations, meanings, and attitudes casting women as unequal to men. Now suppose that Arthur replies,

2 Oh, come on. That’s not true,

and goes on by offering experiential evidence, and even research data, showing that women are generally just as good at driving as men. Arthur here engages in

what we call “denying”: he issues a direct rebuttal to John’s toxic claim, which is deemed false and is rejected on the basis of contrasting evidence.

Direct rebuttals like this perfectly fit the more speech model emerging from Brandeis’s words. Here’s the quote once again, this time in full:

If there be time to expose through discussion the falsehoods and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence.

(Brandeis 1927)

Denying is the prototypical way of “exposing through discussion the falsehoods and fallacies” propagated by toxic speech. It is aimed at stopping their spread, and, provided the counterspeaker disposes of counterevidence and has good argumentative skills, it may indeed force the toxic speaker to concede that they were wrong and prevent other people in the audience from endorsing their views about the targeted group. This is what our counterspeech strategies should ideally aim for.

This strategy, however, has several limits, the first of which is that it is markedly *confrontational*: in directly rebutting what one’s interlocutor has said, one takes an adversarial stance toward them. Sometimes, this is exactly what one should do. Imagine that, during a presidential debate, one of the candidates states that a woman’s place is in the home or that black people are violent. It may be not only appropriate but indeed imperative for the other candidate to openly confront them by forcefully denying their statements. Other times, however, barefaced confrontation may not be the best way to go, all things considered. One may have too few chances to make the toxic speaker drop their claim or convince the audience of the falsity of certain toxic views – and face too high a risk of backlash or retaliation. Denying may be dangerous ground, and particularly so for *targets*: a woman who directly rebuts a sexist statement may face a higher risk of being interpreted as overly sensitive, humorless, or a troublemaker than a man who does so; a black man who openly confronts a racist speaker may risk incurring in particularly harsh forms of retaliation, including physical violence (see, e.g., Rasinski and Czopp 2010; Dickter, Kittel, and Gyurovski 2012). Denying may also be dangerous ground when the counterspeaker, independent of their group membership, is subordinate to the toxic speaker – say, because the toxic speaker is their boss, their manager, or their teacher.

Denying aims at falsifying or disproving the toxic utterance it replies to. But often what is problematic with toxic speech has nothing to do with its content being false (Langton 2018; McGowan 2018). When slurs are hurled as epithets (“You S”) or used in statements aimed at informing the audience of a (supposed) fact (“That S just moved here”), our concern is not primarily with the contribution they give to the utterance’s truth value. Slurs ascribe an inferior status to certain groups of people, function as social mechanisms to push them back “in their place”, and undermine their sense of dignity and assurance of

equal standing (Waldron 2012). This – not falsity – is what concerns us the most. Replying, “That’s not true” in an attempt to counter a slurring utterance would just be missing the point.

Denying may fail as counterspeech even when falsity *is* an important part of the picture. A mere denial is a statement that what one’s interlocutor has explicitly said is not true. As such, it does not address what was implicitly conveyed by the interlocutor’s utterance. When falsehoods are not asserted, but merely implicated or presupposed, denying is not sufficient to rebut them.

Consider a revised version of our “women-just-can’t-drive” example. As in the original version, John, Paul, and Arthur are chitchatting, when they see Sarah in a large SUV. But this time, John says,

3 Wow! That’s huge. No doubt she’ll have her husband park it for her.

John does not explicitly say anything sexist, and yet his utterance implicitly conveys the same sexist content as (1), namely,

4 Women are worse at driving than men.

To retrieve this content, one may reason as follows:

- i John is saying that, undoubtedly, Sarah will have her husband park her SUV for her.
- ii John may be saying so because he thinks that Sarah is worse at driving than her husband.
- iii The context provides no elements to infer Sarah’s or her husband’s actual driving skills.
- iv Sarah is a woman and her husband is a man.
- v Women are stereotyped as worse drivers than men.
- vi John must think and be conveying that Sarah is worse at driving than her husband because of her gender; that is *that women are worse at driving than men*.

Clearly, it is very unlikely that anyone would consciously go through steps (i) to (vi). Typically, we grasp implicit meanings quickly and unreflectively. The above steps offer a rational reconstruction of how one could retrieve (4) – a “conversational implicature”, in technical jargon (Grice 1975) – from (3).

There is widespread agreement in philosophy of language that implicitly communicated content tends to enter the conversational common ground by default unless somebody objects. “Common ground” roughly stands for the set of assumptions that participants mutually share for the purposes of the conversation (Stalnaker 2002).⁴ If uncontested, (3) will sneak the assumption *that women are worse at driving than men* into the conversational common ground. This does not mean that every participant will necessarily end up believing that

women are worse drivers than men. But from then on, the conversation will proceed under the assumption that it is indeed so: (4) will be accepted as true at least for the sake of the conversation.

As a result, the norms in force in that conversation will shift: certain subsequent moves will become appropriate (or “permissible”), whereas others will become inappropriate (or “impermissible”). Making fun of, or joking about, women drivers, for example, will become more contextually appropriate than it was before. Even more than that, John’s move may *encourage* the others to follow suit and play the “gender stereotypes game” as a way of bonding with one another. Conversely, behaviors clashing with what is now common ground will become inappropriate or be discouraged. Suppose that, before seeing Sarah, John, Paul, and Arthur were talking about John’s son and his upcoming driving test. Suppose Paul was about to tell the others how he got his driver license thanks to his girlfriend who taught him how to drive. Once (3) is uttered, and if no one objects, it becomes less easy or is no longer appropriate for him to tell that story.

Implicit content tends to get automatically incorporated into the common ground *but only insofar as nobody objects*. Hearers hold in their hands the power to block the process (Langton 2018). Not any objection will do, though. Imagine that, faced with (3), Arthur replies,

5 I don’t think so.

Even though (5) denies what John explicitly asserts (“No doubt Sarah will have her husband park her SUV for her”), it lets the implicit content *that women are worse at driving than men* pass. A mere denial is indeed compatible with that content, as proven by the fact that (5) could be fleshed out as

(5*) I don’t think so. Sarah is an excellent driver for being a woman!

– which would clearly support, rather than contest, the generic assumption that women are not as good as men at driving.

To counter implicit contents, one should go beyond mere denials and design one’s objection in more sophisticated ways. One such way is what Rae Langton (2018) labeled “blocking”.

Blocking

Paradigmatic blocking is a two-step procedure. The blocker *explicitates* and then *denies* the content implicitly conveyed by their interlocutor. “Explicitation” is a term of art introduced by Marina Sbisa (1999) to name the explicit exposure of implicit content. To counter (3), for example, Arthur might say,

6 Are you assuming she can’t park it because she’s a woman? That’s ridiculous.

Once (6) brings to the surface (or explicitates) the implicit content of (3), it can be easily targeted for denial.

Even though Langton mainly focused on *explicitation-plus-denial*, she acknowledged that blocking is defined by its function:

“Blocking” is a label for a hearer’s resistance to what a speaker, or a speech act, presupposes: “Wait a minute –“ says the hearer, or “Whadd ya mean – *even* George could win?” [...] Blocking interferes with the evolving information taken for granted among participants in a conversation.

(Langton 2018, 148)

This suggests that *any* contribution that prevents some implicit content from automatically becoming common ground will count as a blocking maneuver. So, although *explicitation-plus-denial* has been taken to constitute its paradigmatic form, blocking can come in many guises (Cepollaro n.d.). It may consist of explicitation only, for example. A reply like

(6*) Are you assuming she can’t park it because she’s a woman?

is a blocking maneuver, in that it prevents the common ground from being automatically updated with (4) (“Women are worse at driving than men”). If John still wants (4) to become a shared assumption, he will have to argue for it out in the open.

Sometimes, if one wants to block, it may be enough to stress *that* a certain utterance carries, or may be read as carrying, some implicit content – rather than fully articulating *what* that content is. And, indeed,

7 What are you implying?!

may be just as effective as (6) or (6*) in countering (3).

Blocking is a success term: you cannot block without accomplishing the definitional function of blocking. Turned on its head, this says that, when it comes to blocking, *success* consists in preventing certain implicit contents from entering the common ground by default. This is compatible with such contents eventually making it to the common ground. Suppose that a certain blocking maneuver leads to a discussion surrounding the contested content and that the toxic speaker manages to convince the others of its truth or acceptability. In such a scenario, the contested content eventually becomes a shared assumption. Blocking, however, would still minimally succeed, since the toxic content would not slip into the common ground automatically and unnoticed; that is, *without conversation participants fully realizing it*. Preventing a speaker from smuggling in some controversial content through the “back-door” (as Langton [2018, 152] would put it) and making everyone pay attention to it is per se an important achievement and may serve as counterspeech when that content is toxic.

It is worth pointing out at this point that, according to some, drawing everyone's attention to certain toxic contents is a double-edged sword and risks making paradigmatic instances of blocking backfire.⁵ By unpacking and bringing toxic associations to the surface, paradigmatic blocking may make them more contextually *salient* and thus more cognitively available to participants. This is potentially troubling, because empirical studies suggest that the more an association is cognitively available, the more people are disposed to believe it and to act on it (Lewandowsky et al. 2012). So, increasing the cognitive availability of bigoted associations may further bias people toward bigoted choices and behaviors.

Note that paradigmatic blocking (i.e., explicitation-plus-denial) has denying as one of its constituents, and this makes it *confrontational* in character. Furthermore, and related, paradigmatic blocking can be *face-threatening*: it threatens the "positive face" (or reputation)⁶ of the speaker, who has not *said* anything bigoted and yet is called out for bigotry. As already pointed out, sometimes, openly confronting a toxic speaker and threatening their face is exactly what one should do. When a politician tries to smuggle in some bigoted assumptions, it is of utmost importance that their attempt be brought to light and that they be forced to take responsibility for the toxic contents their words tacitly conveyed. But threatening another person's face may be perceived as aggressive or uncooperative and may lead to backlash and retaliation. Just as denying, blocking may thus be unsuited or unsafe for counterspeakers who are contextually, socially, or institutionally at a disadvantage in comparison to the toxic speaker.

In the next section, we shall look at a counterspeech strategy that operates in a subtler way and thus may come across as less confrontational and less face-threatening. Elsewhere, we called this strategy *bending* (Caponetto and Cepollaro 2022).

Bending

Consider the revised version of our "women-just-can't-drive" example once again. As you will recall, John, Paul, and Arthur are chitchatting, when they see Sarah in a large SUV. John says,

(4) Wow! That's huge. No doubt she'll have her husband park it for her.

Now suppose that Arthur perfectly realizes that John meant to suggest that women are bad drivers. Yet, he replies as if he interpreted John's remark quite differently:

(8) You're right, she should definitely give him parking lessons! He's so bad at parking. But her SUV is new, I doubt she'll trust him with it.

This is a deviant reply to (4). John intended to suggest that Sarah, as a woman, cannot possibly park a large car. Arthur gets it but replies as if John meant that Sarah should let her husband park her SUV as a way for him to practice driving. Arthur *bends* John's move by treating it as conveying a different, less toxic content. Not only will the assumption that women are worse drivers than men fail to enter the common ground by default but if John does not retort, the conversation will proceed under the assumption that women sometimes are *better* drivers than men, as proven by Sarah and her husband's case.

Bending consists in distorting a certain toxic contribution into an innocuous (or at least less toxic) one. It is a form of *acting as if*: the counterspeaker realizes that a given utterance implicitly conveys that *p* (a toxic content) but acts as if they took it to implicitly convey that *q* (an innocuous or less toxic content). In doing so, they prevent *p* from being incorporated into the common ground by default and attempt to make *q* enter the common ground instead – something they will manage to do if the toxic speaker plays along.

Since bending partly relies on toxic speakers playing along, the question as to why they would do so arises. Our answer appeals to social norms of equality. Many ordinary social contexts are governed by norms prescribing people, for example, not to be racist or sexist. Clearly, and problematically, such norms do not preclude people from engaging in everyday racism or sexism. Still, they do pressure people not to do so openly (Saul 2018; Mendelberg 2001). This partly explains why everyday bigotry often (although not always) takes implicit, rather than explicit, forms. Bending distorts an implicitly toxic utterance by making it better aligned with equality norms and gives the toxic speaker a sense that bigotry may not be well received by the audience. By distorting John's contribution, Arthur makes it better aligned with the norm of gender equality, and this may lead John to think that open sexism would not be well received in that context. Faced with (8), John *could* in principle retort and openly commit to the content *that women are bad drivers* ("What?! How could *she* teach *him*? Women just can't drive"). But this would be an open violation of the norm of gender equality – which is generally socially risky, and particularly so after Arthur's countermove.

Bending plays the same function as blocking: it prevents a certain implicit content from entering the common ground by default. In this sense, bending is a form of blocking. However, it is a *distinctive* form of blocking that also attempts to sneak in an "ameliorated content"; that is, a less toxic content than the one conveyed by the speaker's utterance. Qua blocking move, bending succeeds when it fulfills blocking's function. Qua distorting move, it succeeds when, in addition, it manages to make the ameliorated content enter the common ground in place of the toxic content. This is the case when the toxic speaker plays along and does not retort by explicitly asserting the toxic content they were implicitly conveying.

Interestingly, bending operates in a covert manner: the one who bends does not point out *what* was wrong, and not even *that* something was wrong, with the toxic speaker's utterance. This usually makes bending maneuvers *less confrontational* and *less face-threatening* than paradigmatic blocking. By acting as

if John's move did not carry anything sexist, Arthur gives him a chance to tacitly disavow his sexist assumption – to carry on as if he never meant to make it. Arthur gives John an opportunity to preserve his “face”. Bending may thus be preferable to blocking when taking a confrontational stance toward one's interlocutor would be too risky, unwise, or otherwise undesirable. (Conversely, blocking may be preferable to bending when one wants to force the toxic speaker to take responsibility for their sneaky suggestions.)

If carefully crafted, bending maneuvers can also avoid raising the contextual salience of prejudiced associations. In saying,

- (8) You're right, she should definitely give him parking lessons! He's so bad at parking. But her SUV is new, I doubt she'll trust him with it,

Arthur does not make the association between women and poor driving skills any more contextually salient – something he would have done had he opted for paradigmatic blocking instead:

- (6) Are you assuming she can't park it because she's a woman? That's ridiculous.

Note, however, that this virtue of bending is conditional on how bending maneuvers are packaged. Suppose that, instead of (8), Arthur uttered,

- (9) You're right. It's so sad and enraging to see how skillful women are made insecure by a patriarchal society, and think they need to rely on men to carry out basic stuff. I mean, she obviously knows how to park her SUV, she's an excellent driver. And yet, I'm sure she doubts her capabilities and has her husband do it for her.

This reply is an instance of bending: Arthur acts as if John were expressing disappointment at certain gender stereotypes, rather than endorsing them. Yet, it *does* contribute to raising the contextual salience of stereotypes against women (and women drivers in particular).

The characteristic features of bending make it an interesting counterspeech strategy, which may be particularly well suited when the counterspeaker has an interest in not being perceived as too confrontational or in not openly threatening the toxic speaker's face. Bending may also have mitigated salience-raising effects than alternative strategies, although, as we have pointed out, this will depend on how it is crafted.

Saying Nothing

Much of the literature on counterspeech seems to operate under the assumption that remaining silent in the face of a given discursive move is to accept it, at

least for the sake of the conversation. If silence equals acceptance, then saying nothing can never serve as a counterspeech strategy: to counter toxic speech, one necessarily has to speak out against it.

A number of scholars have argued that silence entitles one's audience to presume that one accepts or approves of what has been said⁷ (see, esp., Pettit 1994). While some scholars have explored potential defeaters to the silence–acceptance equivalence (Langton 2007; Goldberg 2018, 2020; Lackey 2018), others have gone as far as to claim that silence can even be expressive of dissent. We will here draw upon recent work by Alessandra Tanesini on eloquent silence as a way of expressing dissent, with the aim of assessing whether silence can, in certain circumstances, constitute a form of counterspeech.⁸

Tanesini (2018, forthcoming) maintains that silence cannot be presumed to communicate acceptance by default. Taking the default (though defeasible) interpretation of silence to be acceptance is to fail to appreciate the distinction between eloquent silences and failures to object. “Eloquent silences” are deliberate silences that are intended to communicate. An eloquent silence is an act and can be a *speech* act (i.e., a communicative means); a failure to object, by contrast, is best thought of as an omission. To substantiate the point, consider the Gricean case of a person who, at a tea party, states that “Mrs. X is an old bag”. Grice (1975, 54) imagines the statement to be followed by an “appalled” silence, after which one interlocutor changes the subject to a discussion about the weather. Unlike Grice, who is primarily interested in the change of subject and how it flouts the conversational maxim of relevance, Tanesini is interested in the silence that precedes it and how it can itself communicate disapproval. Far from being a failure to object, an appalled silence in the face of an inappropriate claim can clearly communicate that “the speaker’s comment should not be dignified with a response” (Tanesini 2018, 118). In a somewhat similar vein, suppose my partner asks me, “Are you still mad at me?” and I deliberately remain silent. My silence is *eloquent*: it communicates an affirmative answer to my partner’s question. By remaining silent, I intend that they believe that I am still mad at them and that they recognize that I have this intention (Tanesini 2018, 114; example adapted from Saville-Troike 1995, 9).

So, not only can silence fail to communicate acceptance, it can even be a way of expressing disapproval or dissent. Interestingly for our purposes, Tanesini (2018) suggested that silence is paradigmatically communicative of dissent when verbal behavior on the part of the silent person would be *expected*. When athletes remain silent as the national anthem plays, their silence clearly communicates disapproval, as they would be expected to sing along. And when a political activist remains silent during an interrogation, their silence is not a mere failure to provide information but communicates their deliberate refusal to do so, as interrogation questions make informative answers expected. In the vocabulary of conversation analysts, “question–answer” is an *adjacency pair*; that is, a two-part exchange in which the first part makes the second relevant

and expected (Schegloff and Sacks 1973; see also Levinson 1983). Other examples of adjacency pairs include “greeting–greeting”, “congratulations–thanks”, “offer–acceptance/refusal”, etc. When silence occurs in place of the second part of an adjacency pair, it *overtly* violates an expectation of verbal behavior. Speakers may exploit this to make their disapproval of something manifest – to communicate a silent implicature of dissent (Tanesini 2018).

This makes room for the possibility that when toxic speech makes verbal behavior of some sort expected, silence can serve as counterspeech. Consider yet another version of our “women-just-can’t-drive” scenario. John, Paul, and Arthur are chitchatting, when they see Sarah getting in a large SUV. John says,

(10) So guys, comments about the gal on her way to smash that car for good?

Suppose (10) is met with silence. No one comments. No one laughs or smiles. Embarrassed, John eventually changes the subject. By remaining silent, Paul and Arthur defeat an expectation of verbal behavior set by John’s question and thereby manage to successfully communicate disapproval. Their silence can be cast as a form of *blocking*: John’s question carries the sexist assumption that women are bad drivers; by keeping silent, Paul and Arthur block its way into the common ground.

Generalizing from this, saying nothing can serve as counterspeech, at least when it defies an expectation of verbal behavior set by the toxic speaker’s move. Clearly, this is not but an initial step into an exploration of silence as counterspeech. Albeit sketchy, however, it interestingly goes against the tide in undermining the assumption that counterspeech requires one to verbally step in.

When silent counterspeech is an option, it may be a particularly well suited one for those who occupy disadvantaged social positions, feel relatively powerless, or have been variously silenced. When speaking up would be an unpromising way to go – for example, because one would not be given the credibility one would deserve – or would be too risky, silence may provide the best (if not the only) shot one has at counterspeaking (Tanesini 2018). Notice, moreover, that silence has no salience-raising effects: since it fails to engage with the toxic content *at all*, it entirely avoids the risk of raising the contextual salience of the prejudicial associations conveyed by the speaker.

Admittedly, though, silence can be employed as counterspeech only in a specific (and perhaps very limited) set of contexts; for example, when a toxic utterance is the first part of an adjacency pair. Furthermore, since the very same action of keeping silent can constitute several different speech acts, eloquent silences seem to be highly vulnerable to be misunderstood, wrongly interpreted as noncommunicative, or distorted into communicative contributions other than those the silent person intended to make (Klieber 2021).

Preemptive Moves

Let us conclude our overview with what may be called “preemptive counterspeech”. Counterspeech moves are prototypically *reactive* or *post hoc*: the counterspeaker par excellence is a speaker who reacts to a given toxic utterance by speaking back against it to remedy or mitigate its harmful effects. Another way to put it is to say that counterspeech is prototypically a *second-turn intervention*: a response to a first-turn contribution conveying, either explicitly or implicitly, something toxic. In the recent philosophical literature, however, the label “counterspeech” is increasingly being used in a broader sense to also include *anticipatory* or *preemptive* moves. The basic idea behind preemptive counterspeech is that one can use language to condition the conversational context *in advance* and in such a way as to make it inhospitable to toxic speech (Tirrell 2018; Lepoutre 2019, 2021). Suppose that the state, through its officials, repeatedly affirms ideals of equality and mutual respect. In doing so, the state would contribute to enacting norms of equality that may render more socially costly, and thus less likely, for citizens to publicly say toxic things.

Preemptive moves of this sort count as counterspeech only where toxic speech is an existing problem. In an ideal world where toxic speech does not exist, promoting ideals of equality would not count as a form of counterspeech. There is thus a sense in which preemptive counterspeech, albeit temporally prior to (potential) toxic utterances, remains a second-turn intervention. Following Tirrell’s (2018) epidemiological metaphor, preemptive counterspeech would be analogous to vaccines: measures introduced *in response* to certain existing diseases, even if they operate *ex ante*.

Maxime Lepoutre (2019, 2021) has recently argued that preemptive moves can alleviate a number of drawbacks often associated with *post hoc* counterspeech. For example, some scholars have expressed the worry that (*post hoc*) counterspeech, even when successful, may ultimately be unable to *undo* the harms of toxic speech: by the time the former comes into play, the latter may have already taken root in a way that cannot be easily reversed (McGowan 2009; Simpson 2013). Preemptive counterspeech is immune to this worry: when successful in conditioning the conversational context, it prevents toxic utterances from even being made.

Lepoutre (2019) also suggested that, *if positively framed*, preemptive counterspeech can avoid salience-raising effects. “Negative counterspeech”, in Lepoutre’s parlance, is counterspeech that repeats a certain toxic content in the process of negating it. Denying and paradigmatic instances of blocking fall under this category. “Positive counterspeech”, by contrast, affirms egalitarian worldviews that implicate the falsity or untenability of certain toxic contents. While repeating a bigoted association to reject it may reinforce its contextual salience, affirming an egalitarian vision of the world that implicates its untenability without repeating it bypasses the problem.

It is no coincidence that our previous example involves state actors. Preemptive moves indeed seem to provide for an especially well-suited form of state-sponsored counterspeech. While state actors rarely find themselves in the position of resisting toxic speech on the spot, they can (and perhaps should) consistently promote education and awareness-raising campaigns, thus serving as preemptive counterspeakers. Lepoutre's (2021) recommendation is that they do so by carefully crafting those campaigns in positive (as opposed to negative) terms.

Preemptive moves are, however, potentially available to ordinary speakers as well. As we saw, toxic speech may shift the norms operative in a given context in harmful ways. As Mary Kate McGowan (2009) has argued, toxic speech may enact oppressive norms by rendering discrimination permissible or (more) appropriate in a given context. More recently, McGowan (2018) suggested that the same norm enactment mechanisms deployed by toxic speech can be used by ordinary speakers to enact egalitarian norms and promote justice. To see how this may be so, an example may help. Until relatively recently, scholars (including philosophers) used to refer to a generic individual by using the male pronoun "he". In this way, they contributed (often unwittingly) to reinforcing the assumption that maleness is the norm.⁹ When feminist scholars started to refer to a generic individual by using the female pronoun "she" or the singular gender-neutral "they" in place of "he", their intervention contributed to the erosion of that normative assumption. Although McGowan does not use this example and does not talk about preemptive moves, we think the case nicely illustrates both what McGowan hints at and preemptive counterspeech. Indeed, not only did feminist scholars turn an androcentric assumption to its head by appropriating the mechanism responsible for its diffusion; they also conditioned the context (e.g., the philosophical arena) by making it gradually less hospitable to moves carrying that assumption.

Conclusion

In this chapter, we went through a number of counterspeech strategies discussed in contemporary philosophy of language. Our investigation makes it clear that the general question "What's the best counterspeech strategy?" is ill formed. When it comes to countering toxic speech, there is no one-size-fits-all solution. As we saw, each of the discussed strategies is promising under certain circumstances and unpromising, and even liable to backfire, under others. Philosophy of language may help us identify the main predictors as to which strategies may be better suited to respond to a given toxic utterance and most likely to succeed in a given conversational context.

Notes

- 1 The definition of *hate speech* is controversial; see, for example, A. Brown (2015) and Anderson and Barnes (2022). See also Lepoutre et al. (2023) for the first corpus-based analysis of the ordinary meaning of hate speech.

- 2 This chapter adopts a philosophy of language angle. For a general overview of the issues that counterspeech raises in philosophy, including moral and political philosophy, see Cepollaro, Lepoutre, and Simpson (2023).
- 3 The importance of contextual factors is equally recognized and discussed by Zollner (chapter 1) in this volume.
- 4 See also Lewis (1979), Langton (2018), and McGowan (2019) for versions of the idea that, insofar as nobody objects, implicit content tends to become common ground by default.
- 5 Lepoutre (2019, 160ff) provides a general discussion of this objection (see also the references therein).
- 6 The *loci classici* for the notion of “face” are Goffman ([1955] 1972) and P. Brown and Levinson (1978, 1987).
- 7 “What is said” captures explicitly conveyed content. This focus on explicit, rather than implicit, communication (and on assertion in particular) is not surprising, since the philosophical debate on the meaning of silence is not mainly concerned with silence in response to toxic speech. We think, however, that some considerations made within this debate can be easily adjusted to address the broader question as to whether silence communicates acceptance of a speaker’s overall contribution; that is, of what they have explicitly and implicitly conveyed.
- 8 The question whether silence equals acceptance is closely tied to a different but related question; that is, whether people have a *duty* to manifestly express their dissent. We will not be concerned with this question here, but see Maitra (2012), Lackey (2018), Langton (2018), and McGowan (2018) for discussion. See also A. Brown (2019) and Saul (2021) on silence and dissent on social media.
- 9 On the false gender-neutrality of “he”, see Moulton (1981).

References

- Anderson, L, and M Barnes. 2022. “Hate Speech.” In *The Stanford Encyclopedia of Philosophy*, edited by EN Zalta and U Nodelman. Stanford, CA: The Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2022/entries/hate-speech/>. Accessed 6 September 2023.
- Brandeis, Louis. 1927. “Concurring Opinion in *Whitney v California*.” http://www.columbia.edu/itc/journalism/j6075/edit/readings/brandeis_concurring1.html. Accessed 6 September 2023.
- Brown, A. 2015. *Hate Speech Law: A Philosophical Examination*. New York: Routledge.
- Brown, A. 2019. “The Meaning of Silence in Cyberspace: The Authority Problem and Online Hate Speech.” In *Free Speech in the Digital Age*, edited by SJ Brison and K Gelber, 207–223. Oxford: Oxford University Press.
- Brown, P, and SC Levinson. 1978. “Universals in Language Usage: Politeness Phenomena.” In *Questions and Politeness: Strategies in Social Interaction*, edited by E Goody, 56–289. Cambridge: Cambridge University Press.
- Brown, P, and SC Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.
- Caponetto L, and B Cepollaro. 2021. “‘Discrimination Preferred’: How Ordinary Verbal Bigotry Harms.” *Australasian Philosophical Review* 5 (2): 189–195.
- Caponetto, L, and B Cepollaro. 2022. “Bending as Counterspeech.” *Ethical Theory & Moral Practice*. doi:10.1007/s10677-022-10334-4
- Carnaghi, A, and A Maass. 2007. “Derogatory Language in Intergroup Context: Are ‘Gay’ and ‘Fag’ Synonymous?” In *Stereotype Dynamics: Language-based Approaches to the Formation, Maintenance, and Transformation of Stereotypes*, edited by Y Kashima, K Fiedler, and P Freytag, 117–134. New York: Lawrence Erlbaum.

- Cepollaro, B.n.d. "A Taxonomy of Blocking Strategies." Unpublished manuscript.
- Cepollaro, B, M Lepoutre, and RM Simpson. 2023. "Counterspeech." *Philosophy Compass* 18 (1): e12890.
- Dickter, CL, JA Kittel, and II Gyurovski. 2012. "Perceptions of Non-Target Confronters in Response to Racist and Heterosexist Remarks." *European Journal of Social Psychology* 42 (1): 112–119.
- Fasoli, F, A Maass, and A Carnaghi. 2015. "Labelling and Discrimination: Do Homophobic Epithets Undermine Fair Distribution of Resources?" *British Journal of Social Psychology* 54 (2): 383–393.
- Fasoli, F, MP Paladino, A Carnaghi, J Jetten, B Bastian, and PG Bain. 2016. "Not 'Just Words': Exposure to Homophobic Epithets Leads to Dehumanizing and Physical Distancing from Gay Men." *European Journal of Social Psychology* 46 (2): 237–248.
- Goffman, E. (1955) 1972. "On Face-Work: An Analysis of Ritual Elements in Social Interaction." In *Communication in Face-to-Face Interaction*, edited by J Laver and S Hutcheson, 319–346. Harmondsworth: Penguin.
- Goldberg, SC. 2018. "Dissent: Ethics and Epistemology." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 40–60. New York: Routledge.
- Goldberg, SC. 2020. *Conversational Pressure: Normativity in Speech Exchanges*. Oxford: Oxford University Press.
- Grice, HP. 1975. "Logic and Conversation." In *Syntax and Semantics: Vol. 3. Speech Acts*, edited by P Cole and JL Morgan, 41–58. New York: Academic Press.
- Klieber, A. 2021. "'Your Silence Speaks Volumes': Silent Implicature and Its Political Significance." PhD thesis, University of Sheffield. <https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.855701>. Accessed 6 September 2023.
- Lackey, J. 2018. "Silence and Objecting." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 82–96. New York: Routledge.
- Langton, R. 2007. "Disenfranchised Silence." In *Common Minds: Themes from the Philosophy of Philip Pettit*, edited by M Smith, R Goodin, and G Brennan, 199–214. Oxford: Oxford University Press.
- Langton, R. 2012. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In *Speech and Harm. Controversies over Free Speech*, edited by I Maitra and MK McGowan, 72–93. Oxford: Oxford University Press.
- Langton, R. 2018. "Blocking as Counter-Speech." In *New Work on Speech Acts*, edited by D Fogal, DW Harris, and M Moss, 144–164. Oxford: Oxford University Press.
- Leader, T, B Mullen, and D Rice. 2009. "Complexity and Valence in Ethnophaulisms and Exclusion of Ethnic Out-Groups: What Puts the 'Hate' into Hate Speech?" *Journal of Personality and Social Psychology* 96 (1): 170–182.
- Lepoutre, M. 2019. "Can More Speech Counter Ignorant Speech?" *Journal of Ethics and Social Philosophy* 16: 155–191.
- Lepoutre, M. 2021. *Democratic Speech in Divided Times*. Oxford: Oxford University Press.
- Lepoutre, M, S Vilar-Lluch, E Borg, and N Hansen. 2023. "What Is Hate speech? The Case for a Corpus Approach." *Criminal Law and Philosophy*. doi:10.1007/s11572-023-09675-7
- Levinson, SC. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Lewandowsky, S, U Ecker, C Seifert, N Schwarz, and J Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13 (3): 106–131.
- Lewis, D. 1979. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8 (3): 339–359.

- Maitra, I. 2012. "Subordinating Speech." In *Speech and Harm. Controversies Over Free Speech*, edited by I Maitra and MK McGowan, 94–120. Oxford: Oxford University Press.
- McGowan, MK. 2009. "Oppressive Speech." *Australasian Journal of Philosophy* 87 (3): 389–407.
- McGowan, MK. 2018. "Responding to Harmful Speech: More Speech, Counter Speech, and the Complexity of Language Use." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 182–199. New York: Routledge.
- McGowan, MK. 2019. *Just Words: On Speech and Hidden Harm*. Oxford: Oxford University Press.
- Mendelberg, T. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Moulton, J. 1981. "The Myth of the Neutral 'Man.'" In *Feminism and Philosophy*, edited by M Vetterling-Braggin, FA Elliston, and J English, 124–237. Lanham, MD: Rowman & Littlefield.
- Mullen, B, and JM Smyth. 2004. "Immigrant Suicide Rates as a Function of Ethnophobias: Hate Speech Predicts Death." *Psychosomatic Medicine* 66 (3): 343–348.
- Rasinski, HM, and AM Czopp. 2010. "The Effect of Target Status on Witnesses' Reactions to Confrontations of Bias." *Basic and Applied Social Psychology* 32 (1): 8–16.
- Pettit, P. 1994. "Enfranchising Silence: An Argument for Freedom of Speech." In *Freedom of Communication*, edited by T Campbell and W Sadurski, 45–55. Aldershot, UK: Dartmouth.
- Saul, J. 2018. "Dogwhistles, Political Manipulation, and Philosophy of Language." In *New Work on Speech Acts*, edited by D Fogal, DW Harris, and M Moss, 360–383. Oxford: Oxford University Press.
- Saul, J. 2021. "Someone Is Wrong on the Internet: Is There an Obligation to Correct False and Oppressive Speech on Social Media?" In *The Epistemology of Deceit in a Postdigital Era: Dupery by Design*, edited by A MacKenzie, J Rose, and I Bhatt, 139–157. Cham: Springer.
- Saville-Troike, M. 1995. "The Place of Silence in an Integrated Theory of Communication." In *Perspectives on Silence*, edited by D Tannen and M Saville-Troike, 2nd ed., 3–18. Norwood, NJ: Ablex.
- Sbisà, M. 1999. "Ideology and the Persuasive Use of Presupposition." In *Language and Ideology*, edited by J Verschueren, vol. 1, 492–509. Antwerp: International Pragmatics Association.
- Schegloff, EA, and H Sacks. 1973. "Opening Up Closings." *Semiotica* 8 (4): 289–327.
- Simpson, RM. 2013. "Un-Ringing the Bell: McGowan on Oppressive Speech and the Asymmetric Pliability of Conversation." *Australasian Journal of Philosophy* 91 (3): 555–575.
- Soral, W, M Bilewicz, and M Winiewski. 2018. "Exposure to Hate Speech Increases Prejudice through Desensitization." *Aggressive Behavior* 44 (2): 136–146.
- Stalnaker, R. 2002. "Common Ground." *Linguistics and Philosophy* 25: 701–721.
- Swim, JK, LL Hyers, LL Cohen, and MJ Ferguson. 2001. "Everyday Sexism: Evidence for Its Incidence, Nature, and Psychological Impact from Three Daily Diary Studies." *Journal of Social Issues* 57: 31–53.
- Tanesini, A. 2018. "Eloquent Silences: Silence and Dissent." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 109–128. New York: Routledge.
- Tanesini, A. forthcoming. "Speech in Non-Ideal Conditions: On Silence and Being Silenced." In *Sbisà on Speech as Action*, edited by L Caponetto and P Labinaz. Cham: Palgrave MacMillan.

- Tirrell, L. 2017. "Toxic Speech: Toward an Epidemiology of Discursive Harm." *Philosophical Topics* 45 (2): 139–161.
- Tirrell, L. 2018. "Toxic Speech: Inoculations and Antidotes." *The Southern Journal of Philosophy* 56: 116–144.
- Waldron, J. 2012. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.

3

SEEING THE FULL PICTURE

The Value of Interdisciplinary Counterspeech Research

Joshua Garland and Catherine Buerger¹

Introduction

The spreading of mis/disinformation and hateful discourse, such as insults, discrimination, or intimidation of individuals (or groups) on the grounds of their ethnicity, gender, religion, or political beliefs, seems to be a ubiquitous problem across many online platforms. Such speech can cause fear of other groups and even potentially incite violence.² The question of how best to address the problem of hate and mis/disinformation online is a challenging one. Solutions must balance the desire to prevent violence and reduce hateful speech while simultaneously protecting civil liberties such as freedom of speech and expression. Often those proposing solutions think first about what governments or tech companies can do to better regulate the speech being shared. For example, one commonly proposed solution is for tech companies to simply censor or remove this type of content from their platform entirely. However, this is problematic for a variety of reasons. There are ethical and legal concerns surrounding infringing upon civil liberties, and there are difficulties in deciding on what should count as hateful discourse. For example, one cannot simply make a list of hateful words, because the meaning of any speech depends both on its content and the context in which it is shared. The same speech may be highly inflammatory in one context and benign in another. Additionally, there are technical challenges in automatically identifying such hateful content and disinformation for removal. As such, before being removed, questionable content often needs to be reviewed by human content reviewers, which is laborious, expensive for platforms, time-consuming, and, in the most egregious cases, psychologically damaging for the reviewers involved. Finally, it is unclear whether censorship is even effective in curbing hate. In fact, Chandrasekharan et al. (2017) showed that censorship may simply move hate to other platforms as well as make the users more toxic on the new platforms.

Counterspeech, the focus of this book, is a potentially promising solution to this difficult sociotechnical problem that sidesteps many of the challenges just discussed such as censorship and the regulation of speech. As summarized in the introduction to this book, counterspeech is defined as assertive responses to hate speech and disinformation, whereby users of online platforms themselves respond to hateful or misleading content to stop it, reduce its consequences, or discourage it (Benesch et al. 2016; Rieger, Schmitt, and Frischlich 2018; Ziegele et al. 2018; Wachs et al. 2019).³ Many, but not all, counterspeech efforts seek to increase civility and deliberation quality of online discussions.⁴ While there have been many excellent studies on counterspeech over the last several years, one important question has remained unanswered: is counterspeech an effective method to curb online hate? Unfortunately, studying the efficacy of counterspeech is extremely challenging. One reason for this is that the notion of “measuring effectiveness” implies that we can do some form of causal inference on this system. Unfortunately, human society is a vastly complex system subject to enumerable outside influences. Real-world events (such as a war or a surge in asylum seekers) can impact how people on social media speak to one another, making it challenging, if not impossible, to isolate the impact of counterspeech.

Because rigorous causal analysis is out of the question due to the complexity of discourse dynamics, it is imperative that we use all of the tools at our disposal to gain any potential insights into the question of effectiveness. As such, it is crucial that we come together as a research community to study this highly interdisciplinary problem from as many angles and with as many unique viewpoints as possible. We argue here that one such potentially fruitful area of interdisciplinary research would be to bridge the gap between the already existing qualitative and quantitative approaches to counterspeech research.

Qualitative studies seek to describe and contextualize complex phenomena and thus can offer unique insight into the human experiences that result from participating in (or being the target of) counterspeech. These studies provide the opportunity for a greater depth of understanding of counterspeech interactions and their effects on society, but they are often hyperfocused on a particular context, and their findings may or may not be generalizable beyond that. Quantitative studies, on the other hand, generally offer breadth and broaden the lens of investigation, which provides insight into the bigger picture through mathematical analysis of large-scale observations of the study system as a whole. However, these studies often lack the detail and ground truth offered by a qualitative approach. While these approaches are clearly complementary, they are rarely performed together. We feel this is a missed opportunity by the counterspeech research community.

In this chapter, we take a first step toward realizing that opportunity by exploring qualitative and quantitative approaches to studying counterspeech and what can be gained by merging the two. We begin with a review of the existing research on counterspeech. Then, we present two case studies of research projects, each undertaken by one of the authors (one qualitative, one

quantitative). One of us (Buerger) is trained as an anthropologist and approaches the topic with an ethnographic mindset. The other (Garland) is an applied mathematician and leans on his expertise in time series analysis, causal inference, natural language processing, and machine learning. These case studies illustrate the specific insights into counterspeech offered by different methodologies and also offer opportunities to reflect on unanswered questions. We conclude with a discussion of what can be learned from reading and collaborating across disciplines.

Literature Review

As mentioned in the Introduction, there is a growing body of scholarship on the topic of counterspeech. Many of these studies focus on gaining insight into whether counterspeech is effective. Though many researchers have tried to answer this question, there are considerable challenges, as we will discuss throughout this chapter. Some of these studies have focused on whether counterspeech is able to effectively change the mind or behavior of a person who has posted hatred online (Schieb and Preuss 2016). Others have focused on the overall discourse within a particular space, asking whether counterspeech has had a favorable impact on it (Álvarez-Benjumea and Winter 2018; Friess, Ziegele, and Heinbach 2021; Garland et al. 2020).

There are also many studies that can be read with an eye to their implications for counterspeech's effectiveness, even if this is not their explicit focus. For example, Berry and Taylor (2017) examined whether the order in which social media comments are ranked had an impact on the comments added to threads. They found that among the users who choose to contribute to the discussion, seeing higher quality comments increased the quality of their subsequent contributions. The authors attributed this effect to the adoption of descriptive norms – social rules based on perceptions of how others are behaving. Studies such as this one have implications for counterspeakers because they help explain some of the mechanisms through which counterspeech may impact discourse.

Masullo et al. (2022) studied the spectator reactions to comments intervening in disagreements between users on the Facebook pages of newspapers. They found that spectators rated intervening comments that used a “high-person-centered response” that acknowledges people’s emotions more favorably than those that used “low-person-centered” speech (Masullo et al. 2022, 494). In another study that examined spectator perception of newspaper comments, Ziegele and Jost (2020) found that “factual responses to uncivil comments increased observers’ perceptions of a deliberative discussion atmosphere” (908), which, in turn, increased their willingness to add their own comments to the thread.

The vast majority of studies on counterspeech have been quantitative. This makes sense, because documenting large-scale trends in discourse requires these methods. There have been a handful of qualitative studies (Abdelkader 2014;

Benesch et al. 2016; Stroud and Cox 2018; Richards and Calvert 2000; Buerger 2021), but they have taken a slightly different focus. These studies generally include case studies of a few counterspeech efforts, attempting to explore the context in which they are embedded and the mechanisms through which they attempt to enact discourse change.

Stroud and Cox (2018), for example, used two case studies to outline a “spectrum of force” of feminist counterspeech, describing how some efforts seek to negatively impact the original misogynistic speaker, whereas others focus on providing support for those targeted by the speech. Benesch et al. (2016) categorized counterspeech examples into a typology of vectors, divided by how the counterspeech functions (one-to-one exchanges, many-to-one, one-to-many, and many-to-many). Buerger’s (2021) ethnography of one large network of counterspeakers examined both the reasons why counterspeakers choose to respond to hateful speech and their strategies for doing so (more on this study below). These qualitative studies provide valuable details about why and how specific counterspeakers seek to change discourse, but they cannot speak to larger trends in the discourse.

A Tale of Two Research Projects

Armed with their own methodological toolboxes, researchers from various disciplines will approach the topic of counterspeech with the ability to ask and answer different questions. In the following section, we provide two case studies of research projects designed to better understand how counterspeech works – one qualitative and quantitative. In each, we describe how our disciplinary backgrounds guided our project framing, methodological approach, and data analysis.

Qualitative Case Study: #jagärhär

One of the largest coordinated counterspeech efforts in the world is the #iamhere network. Originating in Sweden in 2016, the #iamhere network has around 150,000 members responding to hatred through 15 country-level Facebook groups (Sweden, Australia, Bulgaria, Czech Republic, Estonia, France, Finland, Germany, Italy, Norway, Poland, Slovakia, Spain, the UK, and the United States, at the time of writing).

Members of the groups seek out hatred in comment threads of news articles posted on Facebook and then respond together, following a strict set of rules, which includes keeping a respectful and noncondescending tone and never spreading prejudice or rumors. They also “like” each other’s comments, pushing them to the top of comment threads, because Facebook ranks comments on public pages based on interactions (“likes” and replies). This is a vital feature of #iamhere’s model: they make use of Facebook’s display algorithm to amplify their own civil, fact-based comments and bury hateful or xenophobic comments at the bottom of comment threads, making it less likely that others will see them.

In 2018, one of us (Buerger) began an ethnographic study of #jagärhär (the Swedish branch of #iamhere). Anthropologists are trained to study human behavior – how it is influenced by, but also constitutive of, social (and physical) environments. Their methods are often focused on speaking with and observing individuals, but they try to understand these specific data points within a larger landscape of social interaction. Ethnography combines close observation of a particular group or space with in-depth interviews to illuminate aspects of group behavior and culture that are often taken for granted or that may go unspoken among group members.

As more of our daily social interactions have moved online, anthropologists have had to develop new research methodologies to investigate this part of modern human life. Digital ethnography is one such methodology. As with offline ethnographies, anthropologists use ethnographic observation in online spaces to develop an understanding of how the groups they are studying work – what their norms and rhythms are. This observation aids in developing interview questions and in the interpretation of data (Dewalt and Dewalt 2002).

In her study of #jagärhär, Buerger (2021) used digital ethnographic methods to document the daily practices, goals, and decision-making processes of group members. She was particularly interested in the driving factors that had influenced members to begin (and continue) counterspeaking, as well as the strategies they used in their work. The two guiding research questions for the study were (1) how do the external counterspeech actions of group members work to counter hatred? and (2) how do the internal practices of the group keep members engaged? She interviewed 25 members of #jagärhär, which is the original, and largest, branch of #iamhere. Interviews were drawn from a random sample of a list of every member of #jagärhär who had participated (commented or “liked” a post) on the group’s Facebook page over a two-week period ($N = 5580$). Those who agreed were interviewed in English, over Skype or Facebook Messenger, and interviews continued until theoretical saturation was reached (the point where no new themes⁵ emerged from new interviews). In addition to the interviews, Buerger observed the daily practices of a variety of groups in the #iamhere network, joining the #iamhere Facebook groups and regularly visiting their pages, reading updates, and observing the rhythms of the groups. This observation allowed for a deeper understanding of what kind of content was shared and how members responded to it and to each other.

The research produced several notable findings. In response to the question of how the external efforts of #jagärhär work to counter hatred, Buerger found something somewhat surprising: most of the counterspeakers interviewed for the study said they were not hoping that their counterspeech would reach those posting hatred online. In fact, their primary audience was another group entirely – others who might see the speech (and their counterspeech) online – a group one counterspeaker called the “silent crowd”. Members of this group are less visible, since they are primarily reading content instead of adding it. Although measuring whether this strategy works poses a challenge for

researchers, one could imagine a qualitative study where researchers post a call for participants in a comment thread and then interview them about their online behavior and perception of the thread.

Members of #jagårhår described multiple reasons for trying to reach this audience instead of the original speaker. Some said they simply did not think it was a good use of time to speak with those posting hateful comments online, because the chances of changing their behavior were low (a feeling supported by the literature). As one noted, “The trolls will not be affected. They get energy from being debated with. It’s other people that you try to stop from joining in on the hateful speech”.⁶

Instead, many described trying to reach those who have not yet made up their mind about the topic being discussed and therefore could be potentially swayed in different directions by the speech in the comments. Others described the importance of simply documenting their disagreement with the hateful or inaccurate information. “These comment fields can make the impression that most people are hateful; they’re not”, one member stated.⁷

Another posited, “If you have lots of hate comments, maybe you are afraid, and you don’t want to say what you think. But if we are 10–20 people arguing against the hate then I imagine that others will also want to do so”.⁸ By encouraging more people to join their counterspeech efforts, members of #jagårhår hope to decrease the proportion of hatred to counterspeech, thereby positively affecting the overarching tone in the space.

The study also contributed to our understanding of how responding to hatred online affects those doing it. Interviewees reported that taking part in collective counterspeech made them feel braver and more willing to speak up – both online and offline. As one group member stated, “I also think it feels a little easier to give my opinion in different situations offline since I became active in the group. It’s a good school. You get a lot of practice in patience and methods of dealing with different kinds of conversations”.⁹

Quantitative Case Study: Reconquista Germanica/Internet

While qualitative digital ethnography explores the intentions and motivations of those involved in online interactions and other activities, quantitative studies aim to provide insights into the overarching trends and patterns present in a system through mathematical analysis of observations. When performing such a quantitative analysis, especially if that analysis is longitudinal, it is often the case that the quantity of data being analyzed is so large that it cannot be processed by humans alone. For example, in the case study we are about to discuss, tens of millions of tweets and hundreds of thousands of conversations were analyzed. It simply would not have been possible to perform this scale of analysis with humans alone analyzing the text. Instead, quantitative researchers frequently rely on time series analysis, natural language processing, and machine learning algorithms to extract the trends and patterns present in their data.

One of the many fundamental tools that allow for large-scale analysis of text are “classifiers”. Classifiers are mathematical tools that identify which of a set of categories an observation *most likely* belongs to. Typically, classifiers assign a categorical label to an observation as well as an indicator of the classifier’s certainty – for example, a probability – about that label. For example, a trained classifier could discern, with varying degrees of certainty, whether a piece of text is counter, hate, or neutral speech. Once trained, classifiers afford researchers the ability to process and categorize large text corpora – orders of magnitude larger than anything a human could process alone. While classifiers are incredibly useful, obtaining enough labeled data to train a classifier reliably is often a challenge.¹⁰ In this section, we discuss a case study that successfully developed a classifier that was able to identify hate and counterspeech in German political discourse and the insights that were gained from that analysis.

Garland, along with his colleagues (Garland et al. 2020, 2022), performed one of the first large-scale longitudinal quantitative studies of the interactions between hate and counterspeech. To accomplish this, they leveraged a unique situation in Germany where two self-labeling groups engaged in organized online hate and counterspeech on German Twitter. One, called Reconquista Germanica (RG), aimed to spread hate and disinformation about immigrants and promote a radical-right political party (the Alternative für Deutschland) during the 2017 German federal election. The other, called Reconquista Internet (RI), tried to actively resist this discourse using organized counterspeech.¹¹ The presence of these two opposing groups allowed Garland et al. to quantitatively study the impact organized hate and counterspeech had on Germany’s online political discourse during that time. It also afforded them the opportunity to observe and analyze the dynamic interplay between hate and counterspeech over time. By studying these two groups, the primary research question that Garland and his team hoped to gain insight into was whether or not counterspeech was an effective method to curb hate online.

To quantitatively study the impact and dynamics of these two groups, the first step was to train an automated classification system to recognize speech patterns typical of both groups; that is, to classify tweets as either hate or counterspeech. To this end, they collected more than 9 million instances of hate and counterspeech taken from the Twitter timelines of known members of RG and RI. They then used these millions of tweets to train 289 classifiers, each consisting of a fine-tuned doc2vec model (Le and Mikolov, 2014) coupled with a regularized logistic regression function. Each of these classifiers were trained with different parameters and on different subsets of the training corpus providing each classifier a slightly different understanding of the language used by RG and RI. Garland et al. then used an ensemble learning approach where 25 of the top-performing classifiers, which each had a marginally different understanding of the language, examined each tweet and assigned to it a probability of being hate or counterspeech. They then averaged all of these probabilities

together to obtain a final score for each tweet. They then applied a confidence threshold to obtain a final label for that tweet. This effectively allowed them to require the classification system to collectively have some level of confidence before labeling a tweet as either hate or counter speech. With the combined knowledge of each of these 25 independent classifiers, the ensemble classification system was able to automatically and accurately identify speech patterns typical of RG and RI members, achieving accuracy scores in line with state-of-the-art classifiers and on large, balanced, out-of-sample test sets.

Before trusting an automated classification system, however – even one with high accuracy scores – it is often considered good practice to ensure that the automated system agrees with how a human would classify the same example, in this case, whether a given tweet is hate, counter, or neutral speech. Therefore, Garland et al. (2020, 2022) also conducted a crowdsourced study on Mechanical Turk where they hired 28 vetted human judges to manually assess and categorize 5000 tweets (equally spread across the spectrum of hate, counter, and neutral speech) on a scale from 1 to 5, where 1 was “very likely counter speech”, 3 was “neutral”, and 5 was “very likely hate speech”. Each tweet was evaluated by at least two reviewers. Because most tweets were evaluated by a different pair of reviewers, standard interrater reliability scores were not applicable. However, the mean difference in ratings was 0.57. Because this is well within one point on a five-point scale, it suggests reasonable agreement of evaluations by the different raters. Garland et al. then checked to see whether how human judges categorized a given tweet in this subsample of their data was in agreement with how the automated classification system classified that tweet. Overall correlation between the scores of human judges and the scores of the ensemble classification system was $r = 0.94$. When examining hate and counterspeech tweets separately, the correlation was less for counterspeech, $r = 0.75$, compared to $r = 0.93$ for tweets classified as hate.¹² Because these correlations are quite high, the team concluded that their classification system did indeed agree with human judgment.

Since the classifier obtained high accuracy scores and agreed with human judgment, the team felt confident they could use this classification system to accurately analyze and categorize hate, counter, and neutral speech in out-of-sample political discourse at scale. This would allow them to study the potential impact these two groups had on political discourse of the time. However, to do this, they would need a very large sample of political discourse over time. To this end, they collected 131,366 “complete”¹³ political conversations that occurred over four years, between January 2015 (the beginning of the migrant crisis in Europe) and December 2018 on German Twitter.

These conversations grew in response to 22 Twitter accounts of prominent news outlets, politicians, and journalists that were heavily targeted by RG and defended by RI and for which Garland et al. had consistent coverage across the four years of study. Collectively, these conversations contained 1,082,201 replies from 138,979 unique users. As one would expect, the structure, complexity, and

size of each of these conversations varied greatly across this data set. For example, some conversations were very small, receiving only a few or even no replies, whereas others were quite large, receiving as many as 1056 replies. On average, conversations contained 8.24 replies. The structure and complexity of how those replies manifested were also quite interesting and heterogeneous. Take maximum reply depth, for example, or the longest continuous chain of replies in a given conversation. Garland et al. (2020, 2022) observed that most users choose to directly reply to the original post and themselves receive very few if any further replies.¹⁴ In other conversations, very long conversation chains form; the largest such chain was 68 replies long. In aggregate, the conversations had an average maximum reply depth of 2.18. This fascinating and rich data set allowed the researchers to quantitatively observe the potential impacts of both organized hate and counterspeech on German political discourse of that time. To do this, they used the ensemble classification system that was described previously to label each tweet in all 131,366 political conversations as being hate, counter, or neutral speech.

Garland and his colleagues then explored this now labeled conversation data set using several different macro- and microlevel lenses, providing complementary views of the data from different angles. The primary goal of the researchers in Garland et al. (2022) was to gain insights into the effectiveness of counterspeech by measuring how hate and counterspeech interacted with each other on macro and micro levels using the conversation data set just described. On the macro level, they examined the proportion of hate, counter, and neutral speech in these conversations over time to gain insight into changes in the composition of online discourse with and without the presence of organized hate and counterspeech groups. This analysis helped assess the overall deliberation value and civility of discussions as time went on but provided little insight into the microlevel interactions between hate and counterspeakers. For this reason, they also studied microlevel interactions between these two groups. For example, they examined how new replies to a conversation thread shifted subsequent discourse within that conversation. The research team also studied other microlevel measures like shows of support; for example, the number of likes that a hate or counter tweet received.

Taken together, the results built a multifaceted quantitative picture of the dynamics of hate and counterspeech online. In particular, Garland et al. found that across several different indicators, organized counterspeech seemed to contribute to a more balanced and less hateful public discourse. For example, after the formation of the organized counterspeech group (RI), the overall proportion of hate speech decreased and that of counterspeech increased. They also found that the presence of organized counterspeech was predictive of a decrease in the proportion of hate speech and an increase in the proportion of counterspeech in subsequent conversations. After the formation of RI, they saw more support of and engagement with counterspeakers; that is, the number of likes and the length of discussions following counter tweets both increased.

Additionally, with the organized efforts of RI, the results suggested that counterspeech became more effective in steering conversations away from hateful discourse, by providing support to counter tweets and by meeting hateful discourse with more counterspeech. These findings suggest that the presence of supporting peers (in this case, other individuals willing to engage in counterspeech) motivated people to oppose hate speech themselves and to defend its targets. See Garland et al. (2022) for more details about this analysis and the subsequent findings.

There were two primary takeaways from these quantitative studies. In Garland et al. (2020) they showed that hate and counterspeech are detectable and classifiable at a large scale. In Garland et al. (2022), the overall findings suggested that organized hate speech is associated with changes in public discourse and that counterspeech – especially when organized – may help curb hateful rhetoric in online discourse. With this second takeaway it is important to mention a small caveat. Any time one wants to measure the effect of something in a quantitative fashion it is standard to do some form of statistical causal inference on that system. However, when studying human society this is simply not possible due to the enumerable outside influences; for example, terrorist attacks; breaking news; extreme weather; activism; political events like elections, rallies, and speeches; and many others. As such, it is important when studying the dynamics of human society to study the problem using many different lenses to get as clear and broad of a picture as possible.

Unanswered Questions

Both Buerger’s (2021, 2022) and Garland et al.’s (2020, 2022) studies produced valuable findings that advance our understanding of how counterspeech works. But they also leave unanswered questions.

One limitation that both methods share is performing causal inference. Testing for causality – especially as it relates to speech – is inherently tricky. The relationship between speech and action (in our case, counterspeech and changed behavior among those who read it) does not exist in a vacuum. Counterspeech encounters occur amidst innumerable outside influences. This is especially true when attempting to study the impact of speech over time. The more time that passes, the more opportunity exists for cultural and societal factors to change – each of which may also have an impact on discourse norms. For this reason, Garland et al. (2020, 2022) made no causal claims. While the authors performed an important longitudinal study of discourse dynamics between 2015 and 2018, and while many valuable lessons were learned, German society was going through a period of self-reflection during this time. For example, in 2015, 1.1 million refugees entered the country, resulting in intense national conversations about Germany’s ethical responsibilities – and social and economic limitations (Donahue and Delfs 2016). In 2018, there were several large-scale political demonstrations throughout Germany in support of refugees (e.g., the #unteilbar “Indivisible” demonstrations), and in the same year, there were neo-Nazi rallies in Chemnitz where

protestors actively sig-heiled in the streets even in front of police officers – a felony in Germany. Factors such as these without a doubt impacted how internet users spoke about refugees.

Causality is also problematic for qualitative researchers. Ethnographic research and qualitative case studies provide detailed descriptions of the processes of human society and behavior and can therefore contribute to our understanding of why, and how, two things are related (such as speech and behavior). Buerger (2021) used ethnographic methods to zoom in on the individuals involved in counterspeech and examine their motivations and decision-making processes, but these methods are poorly suited for understanding causality at scale.¹⁵ In Buerger’s study, the interviewees reported how they *believed* their method of counterspeech worked to improve online discourse, but ethnography cannot statistically document whether a change has occurred or not.

The study also provided a very detailed picture of why interviewees had chosen to get involved with the collective counterspeech effort, the challenges they face, and how the group’s practices helped support members, but one cannot know how representative these findings are. For example, several members described how counterspeaking as a group made them feel braver. They noted that it was easier for them to write counterspeech in a comment thread when they knew they would not be doing so alone. One could posit from these findings that the presence of counterspeech comments in a thread increases the likelihood that individuals will add additional counterspeech comments (a finding that would challenge some of the research on the “bystander effect”¹⁶), but qualitative methods cannot confirm whether this is the case at scale.

Quantitative methods also have their limitations. Garland et al. (2020, 2022) found that organized counterspeech was associated with less hate and more counterspeech in the future. But the study could not speak to *why* this was occurring. What were the motivations of the individuals taking part in RI? Who were they trying to reach with their counterspeech, and were they successful in doing so? Was the change in discourse documented by the researchers a result of individuals who already shared the beliefs of RI speaking up where they had stayed quiet before, or did RI members convince those in the audience to change their minds about particular issues? These questions are likely better answered using qualitative methodologies such as in-depth interviews.

The Value of Interdisciplinary Counterspeech Research

Through her interviews with counterspeakers, Buerger (2021, 2022) was able to document the beliefs and motivations of individuals responding to hatred online, something that can be difficult to do when working with large data sets. Garland et al. (2020, 2022) showed the dynamics that occur between hate and counterspeech online. Separately, both studies advance counterspeech theory, but they each focus on separate corners of the overarching picture of how counterspeech works. When read together, however, one gets a more complete picture.

For example, both studies demonstrate the power of collective counterspeech campaigns. Garland et al. found that organized counterspeech was better able to shift discourse away from hatred than non-organized counterspeech. The study of the #iamhere network also highlights the value of counterspeaking as a group, showing that members draw emotional support from being part of a group. Confronting xenophobic, racist, and misogynistic content day in and day out is difficult, and interviewees described how emotionally taxing it can be. However, they also described how #jagärhär mitigated some of this. Working as a group helped combat feelings of burnout, likely making the effort more sustainable.

Reading our studies together also provides evidence about one of the pathways through which counterspeech alters discourse. Buerger's study documented that counterspeakers feel more confident challenging hatred when they know they are not responding alone. Garland et al. produced a finding in line with this – that the presence of counterspeech was associated with more counterspeech in the future. Garland et al. show us *what* is happening in the discourse at a large scale (something qualitative methods are not capable of capturing), and Buerger's study provides a possible explanation for *why* that change has occurred.

Conclusion: Bringing It Together

The relationship between speech and behavior is complicated. As is the case with all complex human interactions, the systems in which they exist cannot be grasped by just one discipline. Scholars working in different fields each possess a variety of analytical tools to help them understand the world around them. Each of these tools provides a different lens for looking at the topic being studied and therefore allows researchers to see different aspects of it. Interdisciplinary collaborations allow researchers to see a more complete picture and produce a richer understanding of the nuances and complexities of human behavior.

Reading across disciplines and participating in interdisciplinary collaborations inspire researchers to ask new questions and investigate aspects of the research topic that they have not before. For example, Garland et al.'s (2020, 2022) study of Reconquista Internet and Reconquista Germanica found that after the counterspeech group (RI) was formed, the relative frequency of counterspeech increased. This raises many interesting questions – one of them being whether the existence of a defined enemy (RG) was a motivating factor for people to join and participate in RI. Was there something about the relationship between these two groups specifically that played a role in getting people involved in counterspeech? Qualitative methodologies such as in-depth interviews are perfectly suited for answering such questions.

In-depth qualitative studies on counterspeech can also surface new research directions for quantitative researchers. For example, over the past few years,

Buerger has interviewed over 50 counterspeakers from around the world engaged in a variety of online efforts. By interviewing counterspeakers about their motivations and strategies, their challenges, and their definitions of “success”, Buerger’s (2022) work brings out the varying ways in which counterspeakers think about what it means for their speech to be “effective”. These definitions could be used by quantitative researchers to design studies that are responsive to what counterspeakers are actually trying to do, therefore producing studies that are more useful for practitioners.

While these kinds of multidisciplinary insights can be achieved simply through cross-discipline reading, bringing researchers together at the planning stage of research will produce the most reliable and credible findings. Interdisciplinary research is about collaborating to produce shared knowledge that benefits from the insights of multiple disciplines. It also challenges the blind spots and “taken for granted” knowledge of the researchers’ fields.

An example of one such study that could be done in the future would be a multistage study of a tool designed to help social media users engage in more effective online conversations such as the Analysis and Response Toolkit for Trust¹⁷ or “Seriously”¹⁸ developed by French NGO Renaissance Numérique. The study would begin with an ethnographic investigation of the tool, interviewing users about how they hope to use the tool, observing how they actually engage with it, and then interviewing them again about their experience. This initial qualitative analysis would then inform stage two of the study – a quantitative evaluation of the conversations the counterspeakers engaged in after using the tool. Researchers could then use mathematical tools from natural language processing, machine learning, and time series analysis to quantify the quality and impact of counterspeech produced by these participants while using this tool. As a point of comparison, it would be fascinating to also collect counterspeech conversations where such tools were not used and see whether there is any difference in the quality of counterspeech and the observed impact it had on the conversation. As a final stage of the study, researchers would build on the findings of the first two stages and interview participants about their perspectives of the speech they produced with and without the tool, including how they rate the quality of their speech (and any responses) and whether they believe it had a positive or negative impact on the conversations in which they engaged. These qualitative responses would then be compared with the mathematical assessment of the quality and effectiveness of the speech to determine the level of agreement.

This kind of interdisciplinary research can be a challenge. Researchers must learn to question their own presuppositions and rely less on discipline-specific language. Jargon is often used by academics to signal their membership within a particular discipline community. But it is also an impediment to interdisciplinary collaboration and to reaching broader non-academic audiences. Learning to speak (and write) together can be a major obstacle for interdisciplinary researchers, but doing so will pay off in making the knowledge accessible to a larger audience of practitioners and policymakers.

The field of counterspeech research is young, but it is growing rapidly. As a relatively small community, we have the opportunity to come together and share our insights. By doing so, we can design studies that draw on our varied disciplinary backgrounds to produce results that are reliable, credible, and – most important – useful.

Notes

- 1 Both authors contributed equally to this chapter.
- 2 See www.dangerousspeech.org
- 3 Counterspeech can occur online or offline, but in this chapter, we are focusing our discussion on online counterspeech.
- 4 Some counterspeech efforts seek to shame or punish the original speaker. Others may focus on providing emotional support to the targets of the hateful speech.
- 5 Some of the themes that emerged through this research included the notion of not being alone, the goal of reaching the reading audience instead of the person who has posted hateful speech, and discussions about strategically tagging counterspeech comments with the group’s hashtag.
- 6 Interview with author, 19 September 2019.
- 7 Interview with author, 27 August 2019.
- 8 Interview with author, 18 October 2019.
- 9 Interview with author, 19 September 2019.
- 10 “Labeled data” is simply some form of data that has a known label assigned to it. A tweet combined with the label “counterspeech” is an example of labeled data. In almost all cases, labeled data are necessary to train language models and classifiers to recognize specific categories of data – for example, hate and counterspeech – in out-of-sample data.
- 11 See (Keller and Askanius 2020) for a qualitative analysis and in-depth description of the two groups.
- 12 See (Garland et al. 2020) for more details on this classification system and the evaluation process.
- 13 Here “complete conversation” implies that, at the time of collection, every available reply within a conversation thread was collected. Of course, replies could have been added after collection, but this is highly unlikely due to the age of the conversation threads that were collected. Similarly, some replies could have been deleted prior to collection of the conversation and then those would not be part of the “complete” conversation.
- 14 In her research, Buerger (2021) observed that this is a standard strategy employed by many counterspeakers.
- 15 Although qualitative researchers take a variety of approaches in their engagement with causal studies (Ruffa and Evangelista 2021).
- 16 See Bahador (chapter 4, this volume) for more on the potential effects counterspeech can have on bystanders.
- 17 <https://artt.cs.washington.edu/>
- 18 <https://seriously.org/>

References

- Abdelkader, Engy. 2014. “Savagery in the Subways: Anti-Muslim Ads, the First Amendment, and the Efficacy of Counterspeech.” *Asian American Law Journal* 21: 43. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2264791

- Álvarez-Benjumea, Amalia, and Fabian Winter. 2018. "Normative Change and Culture of Hate: An Experiment in Online Environments." *European Sociological Review* 34 (3): 223–237. doi:10.1093/esr/jcy005
- Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. "Counterspeech on Twitter: A Field Study." Dangerous Speech Project. <https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/>. Accessed 6 September 2023.
- Berry, George, and Sean Taylor. 2017. "Discussion Quality Diffuses in the Digital Public Square." In *Proceedings of the 26th International Conference on World Wide Web*, 1371–1380. International World Wide Web Conferences Steering Committee. <https://arxiv.org/abs/1702.06677>
- Buerger, Catherine. 2021. "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse." *Social Media + Society* 7 (4). doi:10.1177/20563051211063843
- Buerger, Catherine. 2022. "Why They Do It: Counterspeech Theories of Change." Dangerous Speech Project, 26 September 2022. <https://dangerousspeech.org/wp-content/uploads/2022/10/Why-They-Do-It-Counterspeech-Theories-of-Change.pdf>. Accessed 6 September 2023.
- Chandrasekharan, E, U Pavalanathan, A Srinivasan, A Glynn, J Eisenstein, and E Gilbert. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined through Hate Speech." *Proceedings of the ACM on Human-Computer Interaction*, 1–22. New York: Association for Computing Machinery.
- Dewalt, Kathleen M, and Billie R Dewalt. 2002. *Participant Observation: A Guide for Fieldworkers*. Walnut Creek, CA: AltaMira Press.
- Donahue, Patrick, and Arne Delfs. 2016. "Germany Saw 1.1 Million Migrants in 2015 as Debate Intensifies." Bloomberg, 6 January 2016. <https://www.bloomberg.com/news/articles/2016-01-06/germany-says-about-1-1-million-asylum-seekers-arrived-in-2015?leadSource=verify%20wall>. Accessed 6 September 2023.
- Friess, Dennis, Marc Ziegele, and Dominique Heinbach. 2021. "Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions." *Political Communication* 38 (5): 624–646.
- Garland, J, K Ghazi-Zahedi, JG Young, L Hébert-Dufresne, and M Galesic. 2020. "Countering Hate on Social Media: Large Scale Classification of Hate and Counterspeech." In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, edited by Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, 102–112. Cedarville, OH: Association for Computational Linguistics.
- Garland, J, K Ghazi-Zahedi, JG Young, L Hébert-Dufresne, and M Galesic. 2022. "Impact and Dynamics of Hate and Counterspeech Online." *EPJ Data Science*, 11 (1), 3. doi:10.1140/epjds/s13688-021-00314-6
- Keller, Nadine, and Tina Askanius. 2020. "Combating Hate and Trolling with Love and Reason? A Qualitative Analysis of the Discursive Antagonisms between Organized Hate Speech and Counterspeech Online." *SCM Studies in Communication and Media* 9 (4): 540–572.
- Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." *Proceedings of Machine Learning Research* 32 (2): 1188–1196.
- Masullo, Gina M, Marc Ziegele, Martin J Riedl, Pablo Jost, and Teresa K Naab. 2022. "Effects of a High-Person-Centered Response to Commenters Who Disagree on Readers' Positive Attitudes toward a News Outlet's Facebook Page." *Digital Journalism* 10 (3): 493–515.

- Richards, Robert D, and Clay Calvert. 2000. "Counterspeech 2000: A New Look at the Old Remedy for Bad Speech." *BYU Law Review* 2000 (2): 553–586. <https://digitalcommons.law.byu.edu/lawreview/vol2000/iss2/2>. Accessed 6 September 2023.
- Rieger, Diana, Josephine B Schmitt, and Lena Frischlich. 2018. "Hate and Counter-Voices in the Internet: Introduction to the Special Issue." *Studies in Communication and Media* 4: 459–472.
- Ruffa, Chiara, and Matthew Evangelista. 2021. "Searching for a Middle Ground? A Spectrum of Views of Causality in Qualitative Research." *Italian Political Science Review/Rivista Italiana di Scienza Politica* 1–18.
- Schieb, Carla, and Mike Preuss. 2016. "Governing Hate Speech by Means of Counterspeech on Facebook." Paper presented at the 66th *International Communication Association Annual Conference*. Fukuoka: Japan. https://www.researchgate.net/profile/Carla-Schieb/publication/303497937_Governing_hate_speech_by_means_of_counterspeech_on_Facebook/links/5761575408aeada5bc4f783/Governing-hate-speech-by-means-of-counterspeech-on-Facebook.pdf. Accessed 6 September 2023.
- Stroud, Scott R, and William Cox. 2018. "The Varieties of Feminist Counterspeech in the Misogynistic Online World." In *Mediating Misogyny: Gender, Technology, and Harassment*, edited by Jacqueline Ryan Vickery and Tracy Everbach, 293–310. Cham: Palgrave Macmillan. doi:10.1007/978-3-319-72917-6_15.
- Wachs S, MF Wright, R Sittichai, R Singh, R Biswal, E-M Kim, S Yang, M Gámez-Guadix, C Almendros, K Flora, et al. 2019. "Associations between Witnessing and Perpetrating Online Hate in Eight Countries: The Buffering Effects of Problem-Focused Coping." *International Journal of Environmental Research and Public Health* 16 (20): 3992.
- Ziegele, Marc, and Pablo B Jost. 2020. "Not Funny? The Effects of Factual versus Sarcastic Journalistic Responses to Uncivil User Comments." *Communication Research* 47 (6): 891–920. doi:10.1177/0093650216671854
- Ziegele, M, P Jost, M Bormann, and D Heinback. 2018. "Journalistic Counter-Voices in Comment Sections: Patterns, Determinants, and Potential Consequences of Interactive Moderation of Uncivil User Comments." *SCM Studies in Communication and Media* 7 (4): 525–554. doi:10.5771/2192-4007-2018-4-525

PART II

Counterspeech in Context: Media, Culture, and the Legal Framework



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

4

COUNTERSPEECH AS PERSUASION AND MEDIA EFFECTS

Babak Bahador

Introduction

Counterspeech is a communicative response to hate speech that aims to reduce its harmful effects and negative externalities. At its core, counterspeech attempts to persuade various audiences exposed to hate speech to change their opinion and behavior, especially when exposure to such speech influences recipients to cause harm to others, such as through acts of violence. As audience research shows, different audiences will be affected differently through their exposure to hate speech, so it is important to differentiate audiences. This chapter distinguishes four different hate speech and counterspeech audiences – hate groups, violent extremists, the vulnerable, and the public. Within online forums, hate speech and counterspeech attempt to impact not only those directly communicating (speech creators/disseminators and their target audiences) but also a broader network of observers. A fuller analysis of the effects of hate speech therefore needs to consider all online groups potentially impacted. The attempt to influence through communication is not new, of course, and brings hate speech and counterspeech analysis within the domain of the more established persuasion and media effects literature, which dates back a century. This chapter thus examines the concept of counterspeech in relation to the persuasion/media effects body of research. In analyzing this intersection, the chapter examines how different effects may come into play when counterspeech is employed and how they could potentially impact the audiences mentioned differently. The effects that will be considered include spiral of silence, cognitive dissonance, reframing, two-step flow, and herding. The chapter first defines the different audiences of counterspeech and then reviews the tactics to reach them. Finally, it presents the different media effects/persuasion theories and how they could influence the different audiences.

Counterspeech Audiences

Within the context of this chapter, four audiences of counterspeech are examined: hate groups, violent extremists, the vulnerable, and the public. Audiences within this context are groups that are likely to have similar roles in the creation, dissemination, or reception of hate speech and responses to counterspeech. Counterspeech producers, to the degree to which they differentiate audiences, should have different goals for each audience and different expectations regarding what impacts are possible for each. Counterspeech generally has three main goals. The first is to influence opinions and behavior away from hateful beliefs and actions, such as support for violence against groups based on their immutable qualities (Benesch 2014). The second is to reduce the volume of hate speech creation and dissemination and increase the volume of counterspeech, thus changing the balance in former hate-dominated forums (Brown 2016). The third, which ties to the first two, is to reduce the number of people associating with hateful beliefs and practices, especially those who are vulnerable to becoming members of hate groups or violent extremists (Bahador 2021). In examining the different audiences, it is important to point out that they operate on a spectrum of potential harm escalation from the public (least dangerous) to the vulnerable to hate groups to violent extremists (most dangerous). In this regard, individuals can migrate both toward and away from more harmful audiences. Counterspeech is one tool among others that aims to move individuals down the spectrum from more harmful audiences toward, ideally, members of the public who are unlikely to be affected by hate speech. The following section reviews these audiences in the following order – hate groups, violent extremists, the vulnerable, and the public.

Hate groups are defined by the Southern Poverty Law Center (n.d.) as organizations with “beliefs and practices that attack or malign an entire class of people, typically for their immutable characteristics”. Hate groups are most clearly defined by their communication, which is hate speech. In its most blatant form, such speech includes dehumanization, demonization, and incitement to violence, although these more extreme hate speech typologies often emerge after negative speech against groups, which can act as an early warning and foreshadow more extreme speech (Bahador 2020, 2023). When classifying hate groups, analysis often focuses on infamous organizations with formal leadership and membership such as the Ku Klux Klan, Proud Boys, and Oath Keepers (or similar hate groups in other countries outside the United States). However, hate groups operating on the internet are typically loosely organized around an ideology without leadership or a formal structure (Berger 2019; Allen 2020). These groups also tend to have international ties with groups in other countries sharing similar ideologies, allowing them to operate across global “hate highways” (Johnson et al. 2019). In online spaces, clout is measured not by membership or dollars, as is typical offline, but rather by online metrics such as followers or volume of engagement, including likes, comments, and shares. The

internet and social media are often blamed for creating the affordances that enable the spread and growth of hate speech and hate groups. These include providing a forum for bringing like-minded people with hateful beliefs together efficiently, spreading exaggerated and inaccurate grievances that encourage new membership and affiliation, and algorithmic amplification of emotive content such as hate speech, which gets more engagement and is therefore commercially beneficial for platforms.

Hate groups are, at their core, political and aim to remake the political order to align with their beliefs. In the United States and other Western countries, information and disinformation about demographic changes over recent decades has been used to generate a fear of change and promote the idea of a return of society to a past mythical, glorified state, often involving a return to a time of greater perceived morality and religiosity (Gorski and Perry 2022). Political ideologies based on “purifying” society and returning them to past glory, of course, have a long and dangerous history. This is particularly the case when those advocating such views gain political power and attempt to put their idea into practice. In 2017, for example, hate speech by extremist voices on Facebook in Myanmar was identified as a determining factor for enabling an offline environment that supported state violence against the Muslim Rohingya minority, which led to the mass ethnic cleansing of almost 1 million and killing of 9000 civilians (Jakes 2022). The Rohingya, like many despised minority groups around the world targeted for hate, were characterized as foreign invaders who were a threat to the Buddhist majority and frequently framed in dehumanizing terms (Lee 2021).

However, even without state power, online hate groups (including loosely affiliated online ones) are often hotbeds for violent extremism – the second type of audience in this analysis. Violent extremists are almost always a part of hate groups who not only share the ideology but are willing to take it to the next level and put their beliefs into action, often at great personal sacrifice, to bring about their ideal society free from groups they blame for its decline and impurity. Violent extremists can be part of organized groups, such as a terrorist organization (e.g., Islamic State, Lord’s Resistance Army), or “lone wolves” who plan and act on their own. The latter have been responsible for well-known hate-fueled attacks in recent years. In just the United States, the relatively easy access to guns has allowed for many such incidents, including shootings against African Americans in Charleston, South Carolina, and Buffalo, New York; Jews in Pittsburgh, Pennsylvania; Latinos in El Paso, Texas; the LGBTQIAP+ community in Orlando, Florida; and Asian Americans in Atlanta, Georgia.

Hate groups typically claim to represent a much larger group with similar immutable qualities. They typically try to win over members of the larger group through messaging that highlights false or exaggerated grievances against other, often marginalized, groups who are framed as threats to their in-group. To solve the threat and protect their in-group, hate groups advocate various solutions to sanction or entirely remove the out-group, thus freeing their group of

the alleged threat. In recent years, white nationalists in many Western countries have claimed that immigrants and domestic minorities pose great risk to the majority white population by increasing crime, drugs, and other vices. Furthermore, they depict scenarios in which their white or European majority and its culture are replaced through demographic changes, making them into a minority whose superior culture and norms are overtaken by inferior ones (Wilson and Flanagan 2022). This messaging can be particularly convincing to members of the in-group who may be experiencing economic dislocation and personal struggle and are therefore susceptible to easy solutions that promise to alleviate their misfortune by allocate blame onto others. While most members of the larger in-group dismiss such rhetoric and see hate groups as vile, there are nonetheless vulnerable members who are open and even attracted to such ideas and the benefits of association with like-minded people who offer a sense of belonging and shared mission.

The third audience in this chapter is therefore referred to as the vulnerable. The vulnerable are audiences who have immutable qualities similar to hate groups and have the potential to become future supporters. The vulnerable tend to be young, impressionable, and often in their teenage years, going through a transition period and still not fully committed to any ideological beliefs. Many of these young people are highly immersed in digital life and culture, spending significant portions of their time online. It is in online forums where these young “digital natives” often develop an identity and belief system. With an awareness of this context, hate groups use such avenues to reach and aim to recruit the young, offering them kinship and identity (Kamenetz 2018). In recent years in the United States, video games have become a popular means by which to recruit teenage boys and girls, who play video games at rates of 97% and 83%, respectively (Perrin 2018). In these forums, through ads and online chats with strangers playing games, attempts are made to persuade teenagers toward hateful ideology and recruit them into hate groups (Keierleber 2021; Valencia 2021). The vulnerable are likely the most important audiences for counter-speech, because they could potentially go toward embracing a hateful ideology or adopt a strong position against it, depending on their message exposure, the credibility of those communicating such messages, and the norms in the online forums they visit.

The final group in this analysis is the public (or bystanders). These are individuals who are not generally targets of hate speech or counterspeech but are nonetheless often inadvertently exposed to it as they visit the same online forums where such speech occurs. The public are important in the battle for influence because they are usually much larger than the other groups and can therefore influence the general norms of online forums if they decide to engage. The majority of the public will often share the same ideology as counter-speakers but without the motivation to get involved to counter hateful rhetoric unless specifically triggered.

Counterspeech Tactics

Before looking at the potential effects of counterspeech on different audiences, it is important to review how counterspeakers can reach these audiences. Counterspeakers can use a variety of tactics, and selecting which method often depends on which audience is targeted and the forums used by hate speech creators and disseminators, because counterspeakers will generally want to operate in the same forums. Because counterspeech is a response to hate speech, the primary targets of counterspeech are often those who create and disseminate it. These will be members of hate groups, broadly defined, and the subset of this audience that was identified earlier as violent extremists. The latter, of course, will usually be difficult to distinguish from the larger hate group online unless they clearly indicate their willingness to act in an extreme manner, such as using violence. A second, but likely more important, target audience for counterspeech is the vulnerable, because they are the most likely to be influenced away from adopting hateful positions. The final target is the public, whose engagement, as mentioned, can influence the overall norms of online forums if they are motivated to participate.

While hate speech can happen anywhere online, counterspeakers cannot be everywhere and have to consider where their efforts are likely to have the most impact. In this regard, the decision regarding where to counter hate speech depends on which audiences counterspeakers aim to influence. If the primary targets are members of hate groups and violent extremists, it would make sense to go into forums where they concentrate and can potentially be found in large numbers. These might include websites such as 8kun (formerly 8chan) or hate group discussion forums on social media platforms, although many of these have been removed over recent years. If the goal is to prevent the influence and recruitment of vulnerable audiences, public forums where hate groups aim to connect with them and recruit them are prime locations for counterspeakers to operate. This might include monitoring the discussion boards of major newspapers, especially where contentious issues such as immigration are raised. Finally, another approach for counterspeakers to get maximum benefit for their efforts is to consider the intensity of hate speech and when it is likely to spike. This temporal approach is likely to be most relevant during scheduled events-driven periods, such as before, during, and shortly after elections, or following unscheduled shocking events, such as high-profile crimes or acts of terrorism blamed on members of a minority group. Because emotions can spike during such periods, even the broader public's prejudices can be tapped into by populists claiming to alleviate fears through collective punishment of groups that become guilty by association. In such scenarios, counterspeakers can try and provide counternarratives that ask the populace to look at their core societal values and how short-term actions contradicting them could harm their society more than any short-term perceived benefits. For election periods, one example of a successful counterspeech effort was that of NGO Sisi Ni Amani, which was

set up to prevent post-election violence in Kenya in 2013 through peace text messages to challenge false rumors that had played an important role in the post-election violence of 2007–8 that claimed 1300 lives (Martin-Shields 2013; Shah and Brown 2014).

In conducting counterspeech online, Wright et al. (2017) identified four tactics. The first is one-to-one, in which a single counterspeaker reaches a single individual espousing hate. The second method is one-to-many, in which one counterspeaker reaches a group sharing hate speech. This tactic might be common, for example, in some public forums where a counterspeaker challenges the prevailing hateful sentiment. This approach can also involve a counterspeaker reaching a group using a hateful hashtag. The third tactic is many-to-one, in which many respond to one message, especially when that message has gone viral. The fourth and final tactic is many-to-many, in which many counterspeakers engage many other speakers, with some portion likely including hateful speakers. This can begin with a broad challenge but can ideally lead into constructive conversation over a timely or controversial topic. In one such example, the hashtag #KillAllMuslims trended on Twitter as a magnet for hate but was then challenged by counterspeakers who reacted collectively to eventually take it over and change its focus away from hate (Wright et al. 2017).

Counterspeech Effects

In examining online counterspeech and its possible effects, it is important to differentiate the methods used to derive findings and reach conclusions. In this regard, at least three methods are worth mentioning. The first includes naturalist studies and involves observing counterspeech as it occurs organically online without any intervention. In such studies, researchers observe the online behavior of different individuals and groups (also referred to as ethnographic research when groups are observed). Findings derived from gathering and analyzing data from actual online communication are likely to be the most reflective of reality, because they are not based on any outside interference. While naturalist and ethnographic research is often qualitative and conducted on a smaller scale in offline settings, online social media data streams, in conjunction with content analysis, allow for the possibility of representative data sets and generalization. These studies, however, have limitations because researchers will still have to interpret beliefs and motivations from text and will not be able to ask subjects questions or monitor offline behavior.

To overcome some of these shortcomings, but also introduce new ones, researchers often use experimental methods to better control variables and get deeper insights from participants. In such settings, different audiences are recruited and know that they are participating in research in a lab setting, thus limiting the authenticity of the exercise. However, the ability to manipulate variables, gather information on beliefs and motivations, and better ascertain information on possible offline actions offers a deeper understanding about

potential effects. Finally, a third approach worth highlighting involves observing interventions by activists working to counter real hate speech online. In these observations by activists such as the group #ichbinhier and #jagärhär (German and Swedish for #Iamhere), researchers observe coordinated counterspeech campaigns that try to reduce hate speech and potential offline harms from it (Porten-Cheé, Kunst, and Emmer 2020; Buerger 2021; see also chapter 3 in this volume). Studying such efforts is authentic because both activists and those producing hate speech are in an actual online setting and not a purely artificial one such as those in experimental studies. Furthermore, this approach allows researchers to conduct interviews with counterspeakers who can share their deeper insights on their efforts.

Counterspeech efforts are conducted because those engaging in them assume that opinions, attitudes, beliefs, and behavior can be changed; otherwise, there would be little point in such efforts. It is worth mentioning, however, that counterspeech, like all attempts at persuasion, is often likely to be ineffective, especially if the targets are those espousing or sharing hate speech. This is because media effects and persuasion research show that it is very difficult to change entrenched views once they are established. Those who make and disseminate hate, such as members of hate groups or adherents to such ideologies, are likely to be committed to their prejudiced and hateful beliefs. However, while rare, there are instances when people espousing hate have been persuaded to change their views and beliefs. A prominent convert was the granddaughter of the founder of Westboro Baptist church, who was the church's Twitter spokeswoman and used the forum to initially share hateful views, especially against gays. However, through exchanges of views over time, she eventually changed her positions once she saw the humanity in her rivals and ultimately became an advocate against her previous hateful and extreme positions (Phelps-Roper 2019).

Furthermore, effects also occur not just by changing minds but also in changing the nature of the arena where hate might be spread, making prominent online spaces formerly festering with hate safe from it, thus removing potential recruitment grounds for the vulnerable. In the following section of this chapter, five key findings from the media effects and persuasion literature that relate to counterspeech and its effects, broadly defined, are outlined. These are the spiral of silence, cognitive dissonance, reframing (rehumanizing), two-step flow, and herding. After each is explained, analysis is provided on how each can also potentially impact the different audiences presented earlier in this chapter and how some effects are more relevant to some audiences than others.

Spiral of Silence

People in the offline world generally do not share their views if they perceive them to be unpopular, out of fear of social isolation, but do share them if they think they are popular. To this end, they seek social cues to understand what is

unpopular, such as criticism and ridicule. This phenomenon has been termed the “spiral of silence” (Noelle-Neumann 1974). Similarly, it appears that most people who engage in hate speech online stop participating when they think that the online sentiment is changing away from hate and their views are becoming the minority view (Miškolci, Kováčová, and Rigová 2020). People generally like to participate in online forums when their views are confirmed and reinforced by other like-minded people. Social media provides the affordances by which those holding minority views, such as hating others based on immutable qualities, could become the majority within specific online spaces and thus feel emboldened to share their prejudiced views without concern for backlash.

While it is difficult to counter hate in all online spaces, counterspeakers can nonetheless monitor and engage in prominent online public spaces, such as the comments section of major newspapers on social media platforms, especially when articles that could draw hateful comments, such as those related to immigration and crime, are published. When counterspeakers coordinate their activities in large numbers in such spaces, they can change the online norm, or what is generally considered acceptable and unacceptable, and turn hateful speakers into the minority representing unpopular views, thus incentivizing many of them to disengage, turning the online norm in that space against hate.

This effect is most likely to impact hate groups and violent extremists, who will become less active in more prominent and potentially influential online spaces. While they may migrate to other smaller forums or closed messaging apps, their ability to influence is nonetheless diminished. The changing of online norms is also likely to reduce the likelihood of vulnerable audiences being exposed to hate speech, thus reducing the odds of their recruitment into hate groups, and likely to increase the likelihood of them adopting positions against hate as they adopt stronger opinions and beliefs. For the public, coming across more frequent comments against hate could give them motivation to participate and support such positions, because most of the public will hold such views but only exercise them when they feel safe and inspired.

Cognitive Dissonance

Individuals feel comfortable when their opinions and beliefs are consistent and uncomfortable when they hold contradictory positions, or what is called “cognitive dissonance” (Festinger 1957). This is alleviated in various ways, such as rationalization and avoidance, but also sometimes by changing one’s previous views on particular issues to create cognitive consistency. To maintain prejudice against an entire group, one must believe that all members of the group are the same and think and act the same, thus making them all guilty of any wrongdoing by one. However, this assumption is false, irrational, and easy to refute with evidence. This is unpacked when those espousing group-based prejudice and hate are asked whether the assumptions they make about other groups

apply to their own group, which they usually challenge and allocate blame to just the guilty individual. In one study that demonstrated this point using experimental research, attempts were made to use different interventions to reduce collective blame and hostility against Muslims by Americans (Bruneau, Kteily, and Falk 2018). Out of nine different intervention types and one control group, researchers found that the most effective intervention involved exposing in-group hypocrisy. First, participants were asked whether Muslims should be collectively blamed for terrorist acts by individual Muslims. Then they were asked whether the same collective blame should be applied to white American Christians for a similar act. While some respondents who blamed Muslims collectively did not blame American Christians collectively, others changed their minds when presented with this inconsistency and, as a result, reduced their prejudice. This was likely because of the cognitive dissonance of holding two contradictory views involving their in-group versus an out-group, which could be alleviated through changing existing views on the out-group.

When monitoring hate speech, it is common for hate groups and violent extremists to conflate the actions of individuals with the group and assign the latter with collective responsibility and blame, while also proposing harm against them, including discrimination and violence. Counterspeakers can respond to such assertions by simply asking whether members of the in-group should collectively be subjected to similar treatment for negative individual acts. Such an approach, if delivered in the form of constructive dialogue, can potentially influence all audiences. While some hateful speakers might rationalize the contradiction and hold their views, a small minority may reduce their prejudiced views and assign collective blame. Furthermore, such a dialogue could certainly influence the vulnerable to move toward less prejudiced views and encourage the public to engage in supporting counterspeech of their own.

Reframing (Rehumanizing)

Because any issue can involve thousands of possible facts, framing is a technique by which content creators highlight some aspects of perceived reality and ignore others to influence audiences in their interpretation. Frames define problems and causes and offer moral evaluation and treatment recommendations (Entman 1993). When hate speech is employed, a particular kind of framing is used in which members of a group with immutable qualities such as gender, nationality, or race are stripped of their individuality and become a category with negative attributes. In essence, de-individualization is a form of dehumanization, because humans are unique with their own personalities and features, but dehumanized categories are represented by a negative stereotype. A key goal of counterspeakers is to rehumanize and thus re-individualize again by challenging the negative collective stereotype. This can be done by referring to members of the group as individuals with unique personalities, hopes, and aspirations. This is a form of reframing that directly challenges the framing of hate speech (Bahador 2012).

If successfully executed, reframing by re-individualizing can undermine a key tactic of hate groups and violent extremists, which hinges on a negative collective stereotype of other groups who they can collectively blame and sanction. This will likely weaken the drive of hate groups and violent extremists, who, especially in the latter case, carry out violence by convincing themselves that they are not harming humans but a sub- or superhuman enemy image. Once that image is deconstructed, the motivation to act is weakened (Bahador 2015). Additionally, vulnerable audiences who hold grievances will not be able to allocate blame to other groups when such groups are no longer seen as a collective negative stereotype but rather as a set of individuals like anyone else.

Two-Step Flow

The two-step flow theory posits that media influence occurs through opinion leaders (Lazarsfeld, Berelson, and Gaudet 1944; Katz 1957). These include political leaders, religious leaders, celebrities, and what today are known as social media influencers. While this notion has been backed by empirical evidence in recent decades, the idea dates back at least to Aristotle, who argued that “ethos”, or the credibility of the speaker, was critical for effective rhetoric and persuasion. To create effective online counterspeech, research continues to show that source credibility matters for persuasion (Briggs and Feve 2013; Munger 2017). This means that counterspeech creators need to take the time to understand their audience and local context and know the sources considered credible and not just counter messages from their own point of view, which may be alien or even offensive to the target audience. Ideally, they can get credible sources to communicate directly to the target audience. For example, in potentially contentious elections, political leaders and media personalities from different parties and factions have collectively encouraged their followers to avoid hate speech and violence, and this is believed to have played an important role in preventing election violence. If the credible sources are not available or cannot be organized in time, however, counterspeakers can use past statements by them or appeal to audiences through citing revered texts such as holy books or constitutions, which they audience may hold in high regard.

Counterspeech with the help of credible sources has shown to be effective with hate groups. For example, in one study, it was shown that white nationalists responded more when the speaker was a conservative who had more similarities to them (Briggs and Feve 2013). When appealing to vulnerable youth against hateful rhetoric, counterspeakers have opportunities to use celebrities and social media influencers who appeal to their demographic and also target them in social media platforms and other spaces they communicate, such as through video game chats with other players. The key, again, is to know your audience, where they spend time online, and who they consider credible.

Herding

Earlier in this section, it was argued that counterspeech efforts, especially when coordinated, can begin to change online norms. This is partially spurred by the reduction of hate speech from those formerly advocating it due to the spiral of silence effect, which deters them from participating. However, the other side of this coin involves increasing anti-hate messaging by getting more participants, especially from the public (or bystanders). This is important because research shows that a lack of a response to hate speech gives targets the impression that bystanders agree with the sentiment (Leonhard et al. 2018). Bystanders often get involved in countering hate if they believe the hate speech is threatening to the targets. They also tend to get more involved with defending some groups over others, based on the perception of incivility toward them (Obermaier, Schmid, and Rieger 2023).

However, just the act of counterspeech is enough for members of the public to get involved. In a number of studies, the introduction of counterspeech motivated others who were not part of any initial effort to join and add comments aligned with the counterspeaker positions (Foxman and Wolf 2013; Costello, Hawdon, and Cross 2017). This is because those who enter online spaces are influenced by the norms that are in effect (Cheng et al. 2017; Kwon and Gruzd 2017; Molina and Jennings 2018). Kramer, Guillory, and Hancock (2014) referred to this as emotional contagion and described how the volume of positive versus negative sentiment leads to posts in the same direction by those entering the forum. The effect by which online norms influence sentiment and spur action can be referred to as a herding effect. “Herding” is defined as a phenomenon in which individuals make decisions based on imitating and following others versus deciding independently based on their own information (Baddeley 2010).

In some ways, herding is the opposite of the spiral of silence. It promotes engagement, whereas the spiral of silence discourages it, both based on the perceived views of others. Herding likely will have little impact on hate groups and violent extremists because it is unlikely for them to change their views and follow the online sentiment away from hate. However, it is likely to have the most impact on the vulnerable, who may be sitting on the fence on choosing their disposition toward hate. While they may share the same grievances as hate groups, as mentioned earlier, they have also likely been taught that group-based hate and prejudice are wrong. When they come onto online forums, the direction of those forums will likely have a notable influence on youth, so it is in such spaces that counterspeakers can have the greatest impact.

Conclusion

This chapter has attempted to connect the relatively new and small body of research on countering hate speech, or counterspeech, with the much older and larger literature on media effects and persuasion. Counterspeech, at its core,

aims to change hearts and minds and, through different media effects, has different impacts on different audiences. It is likely the most effective against hate groups and violent extremists when it can dissuade them from participating in online forums due to an online variation of the spiral of silence effect, in which those who hold unpopular views are likely to avoid espousing them. When it comes to the vulnerable and public (or bystanders), counterspeech is likely the most effective when it can be part of a herding effect, in which the overall sentiment and online norms are shifting away from hate. In such scenarios, the vulnerable, who have not yet committed to a particular ideology, are more likely to adopt views against hate, whereas the public observing debates online is more likely to also join counterspeech efforts on the side fighting hate, thus bolstering the effort.

While a number of studies on counterspeech have directly and indirectly considered how the persuasion and media effects body of research is important for understanding the effectiveness of counterspeech, this chapter has placed the literature more centrally and used it to identify five distinct counterspeech “effects” buckets. This attempt to synthesize the counterspeech literature within the context of persuasion and media effects literature, in fact, is the main contribution of the chapter to the counterspeech literature. Furthermore, the appreciation of this larger literature can help us understand the limits of counterspeech, because many of the debates on how effective counterspeech is have already been considered for decades in the persuasion and media effects literature.

In considering the direction of future counterspeech research, the larger media effects literature and its historic evolution are insightful. Concerns with media effects emerged in the 1920s and 1930s, when there was a general fear that propaganda could be used to control the masses. Observers saw how autocratic regimes at the time, in particular, seemed to persuade and manipulate their people with the use of these new communication techniques and feared an inability to stop them. The so-called hypodermic needle model, in fact, assumed that information in the skillful hands of a propagandist could be injected into the masses, who would then accept the message and be vulnerable to manipulation.

With time and more sophisticated research from the 1940s onward, researchers realized the limits of any mass effect, largely debunking the notion. Instead, researchers found that media effects were highly context and individual specific and that predispositions were key to understanding the likely effect of media messages on different groups and individuals. In much the same way, there could be an early hysteria at this time regarding the role of hate speech on the internet and social media and a belief among some observers that mass response is indeed possible. This exaggerated belief in the power of messages and their ability to stimulate a mass response, to a lesser degree, could also exist among counterspeakers. If the history of media effect research is any guide, future research on the power of hate speech and, to a lesser degree, the effectiveness of counterspeech should further explore the importance of individual preferences and predispositions and appreciate the limited ways in which any effects are likely to take place.

References

- Allen, JR. 2020. "White-Supremacist Violence Is Terrorism." *The Atlantic*, 24 February. <https://www.theatlantic.com/ideas/archive/2020/02/white-supremacist-violence-terrorism/606964/>. Accessed 10 September 2023.
- Baddeley, M. 2010. "Herding, Social Influence and Economic Decision-Making: Social-Psychological Neuroscientific Analysis." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365 (1538): 281–290.
- Bahador, B. 2012. "Rehumanizing Enemy Images: Media Framing from War to Peace." In *Forming a Culture of Peace: Reframing Narratives of Intergroup Relations, Equity and Justice*, edited by KV Korostelina, 195–211. Cham: Palgrave Macmillan.
- Bahador, B. 2015. "The Media and Deconstruction of the Enemy Image." In *Communication and Peace: Mapping an Emerging Field*, edited by V Hawkins and L Hoffmann, 120–132. New York: Routledge.
- Bahador, B. 2020. *Classifying and Identifying the Intensity of Hate Speech*. Social Science Research Council Items. <https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/>. Accessed 1 February 2023.
- Bahador, B. 2021. "Countering Hate Speech Online." In *The Routledge Companion to Media Misinformation and Populism*, edited by H Tumber and S Waisbord, 517–528. New York: Routledge.
- Bahador, B. 2023. "Monitoring Hate Speech and the Limits of Current Definition." In *Challenges and Perspectives of Hate Speech Research*, edited by C Strippel, S Paasch-Colberg, M Emmer, and J Trebbe, 291–298. Berlin: Boehland & Schremmer. <https://www.ssoar.info/ssoar/handle/document/86421>. Accessed 10 May 2023.
- Benesch, S. 2014. "Countering Dangerous Speech: New Ideas for Genocide Prevention." doi:10.2139/ssrn.3686876
- Berger, JM. 2019. "The Strategy of Violent White Supremacy Is Evolving." *The Atlantic*, 7 August. <https://www.theatlantic.com/ideas/archive/2019/08/the-new-strategy-of-violent-white-supremacy/595648/>. Accessed 10 September 2023.
- Briggs, R, and S Feve. 2013. *Countering the Appeal of Online Extremism*. Policy Briefing. Institute of Strategic Dialogue. https://www.dhs.gov/sites/default/files/publications/Countering%20the%20Appeal%20of%20Extremism%20Online_1.pdf. Accessed 1 February 2023.
- Brown, R. 2016. *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech*. <https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf>. Accessed 1 February 2023.
- Bruneau, E, N Kteily, and E Falk. 2018. "Interventions Highlighting Hypocrisy Reduce Collective Blame of Muslims for Individual Acts of Violence and Assuage Anti-Muslim Hostility." *Personality and Social Psychology Bulletin* 44 (3): 430–448.
- Buerger, C. 2021. "#Iamhere: Collective Counterspeech and the Quest to Improve Online Discourse." *Social Media + Society* 7 (4). doi:10.1177/205630512111063843
- Cheng, J, M Bernstein, C Danescu-Niculescu-Mizil, and J Leskovec. 2017. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW '17*, 1217–1230. New York: Association for Computing Machinery.
- Costello, M, J Hawdon, and A Cross. 2017. "Virtually Standing Up or Standing By? Correlates of Enacting Social Control Online." *International Journal of Criminology and Sociology* 6: 16–28.

- Entman, RM. 1993. "Framing toward Clarification of a Fractured Paradigm." *Journal of Communication* 43 (4): 51–58. <https://academic.oup.com/joc/article-abstract/43/4/51/4160153?redirectedFrom=fulltext>
- Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.
- Foxman, AH, and C Wolf. 2013. *Viral Hate: Containing Its Spread on the Internet*. New York: Palgrave Macmillan.
- Gorski, PS, and SM Perry. 2022. *The Flag and the Cross: White Christian Nationalism and the Threat to American Democracy*. New York: Oxford University Press.
- Jakes, L. 2022. "Myanmar's Military Committed Genocide against Rohingya, U.S. Says." *New York Times*, 21 March. <https://www.nytimes.com/2022/03/21/us/politics/myanmar-r-genocide-biden.html>. Accessed 1 February 2023.
- Johnson, N, R Leahy, N, Johnson Restrepo, N, Velasquez, M, Zheng, P, Manrique, P, Devkota, and S Wuchty. 2019. "Hidden Resilience and Adaptive Dynamics of the Global Online Hate Ecology." *Nature* 573: 261–265. <https://www.nature.com/articles/s41586-019-1494-7>. Accessed 1 February 2023.
- Kamenetz, A. 2018. "Right-Wing Hate Groups Are Recruiting Video Gamers." NPR, 5 November. <https://www.npr.org/2018/11/05/660642531/right-wing-hate-groups-are-recruiting-video-gamers>. Accessed 6 September 2023.
- Katz, E. 1957. "The Two-Step Flow of Communication: An Up-to-Date Report on a Hypothesis." *Public Opinion Quarterly* 21 (1): 61–78.
- Keierleber, M. 2021. *How White Extremists Teach Kids to Hate*. The 74. <https://www.the74million.org/article/where-hate-is-normalized-how-white-extremists-use-online-gaming-communities-popular-among-teens-to-recruit-culture-warriors/>. Accessed 1 February 2023.
- Kramer, ADI, JE Guillory, and JT Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences of the United States of America* 111 (24): 8788–8790. doi:10.1073/pnas.1320040111
- Kwon, KH, and A Gruzd. 2017. "Is Offensive Commenting Contagious Online? Examining Public vs Interpersonal Swearing in Response to Donald Trump's YouTube Campaign Videos." *Internet Research* 27 (4): 991–1010.
- Lazarsfeld, PF, B Berelson, and H Gaudet. 1944. *The People's Choice: How the Voter Makes Up His Mind in the Presidential Campaign*. New York: Columbia University Press.
- Lee, R. 2021. *Myanmar's Rohingya Genocide: Identity, History and Hate Speech*. London: Bloomsbury.
- Leonhard, L, C Rueß, M Obermaier, and C Reinemann. 2018. "Perceiving Threat and Feeling Responsible How Severity of Hate Speech, Number of Bystanders, and Prior Reactions of Others Affect Bystanders' Intention to Counterargue against Hate Speech on Facebook." *Studies in Communication and Media* 7: 555–579.
- Martin-Shields, C. 2013. "Inter-Ethnic Cooperation Revisited: Why Mobile Phones Can Help Prevent Discrete Events of Violence, Using the Kenyan Case Study." *Stability: International Journal of Security & Development* 2 (3): 1–13.
- Miškolci, J, L Kováčová, and E Rigová. 2020. "Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia." *Social Science Computer Review* 38 (2). doi:10.1177/0894439318791786?journalCode=ssce
- Molina, RG, and FJ Jennings. 2018. "The Role of Civility and Metacommunication in Facebook Discussions." *Communication Studies* 69 (1): 42–66.
- Munger, K. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39 (3): 629–649.

- Noelle-Neumann, E. 1974. "The Spiral of Silence A Theory of Public Opinion." *Journal of Communication* 24 (2): 43–51.
- Obermaier, M, UK Schmid, and D Rieger. 2023. "Too Civil to Care? How Online Hate Speech against Different Social Groups Affects Bystander Intervention." *European Journal of Criminology* 1–17.
- Perrin, A. 2018. *5 Facts about Americans and Video Games*. Pew Research Center. <https://www.pewresearch.org/fact-tank/2018/09/17/5-facts-about-americans-and-video-games/>. Accessed 1 February 2023.
- Phelps-Roper, M. 2019. *Unfollow: A Memoir of Loving and Leaving the Westboro Baptist Church*. New York: Farrar, Straus and Giroux.
- Porten-Che , P, M Kunst, and M Emmer. 2020. "Online Civil Intervention: A New Form of Political Participation under Conditions of a Disruptive Online Discourse." *International Journal of Communication* 14: 514–534.
- Shah, S, and R Brown. 2014. "Programming for Peace: Sisi Ni Amani Kenya and the 2013 Election." Report, University of Pennsylvania. <https://repository.upenn.edu/handle/20.500.14332/1302>. Accessed 6 September 2023.
- Southern Poverty Law Center. n.d. *What Is a Hate Group?*<https://www.splcenter.org/20220216/frequently-asked-questions-about-hate-and-antigovernment-groups#hate-group>. Accessed 1 February 2023.
- Valencia, M. 2021. "A Hate Group Targeted My Kid Online." *New York Times*, 17 October. <https://www.nytimes.com/2021/09/08/parenting/online-hate-groups-kids.html>. Accessed 1 February 2023.
- Wilson, J, and A Flanagan. 2022. "The Racist 'Great Replacement' Theory Explained." *Hate Watch*. Southern Poverty Law Center. 17 May. <https://www.splcenter.org/hatewatch/2022/05/17/racist-great-replacement-conspiracy-theory-explained>. Accessed 1 February 2023.
- Wright, Lucas, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. "Vectors for Counterspeech on Twitter." In *Proceedings of the First Workshop on Abusive Language Online*, edited by Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, 57–62. Vancouver: Association for Computational Linguistics.

5

REIMAGINING THE CURRENT REGULATORY FRAMEWORK TO ONLINE HATE SPEECH

Why Making Way for Alternative Methods Is Paramount for Free Speech¹

Jacob Mchangama and Natalie Alkiviadou

Introduction

Freedom of expression and internet freedom have been in global decline for over a decade. The global free speech recession has reached a critical state (Freedom House 2021b; V-Dem Institute 2021). The 2022 V-Dem Democracy Report concluded that a record 35 countries are threatening freedom of expression, as opposed to 5 in 2011 (V-Dem Institute 2022). Global internet freedom declined for a 12th consecutive year in 2022, while the Committee to Protect Journalists (CPJ) has documented an increase in the imprisonment of journalists, including a marked increase in the number of journalists imprisoned for spreading “false information” (Freedom House 2022b; CPJ 2023). The pandemic has accelerated the free speech recession. In 2020 the Future of Free Speech Project (FFS) recorded at least 70 global censorship measures formally targeted at COVID-19 disinformation but frequently targeting political dissent and critical media (Mchangama and Alkiviadou 2020b). A 2022 UNESCO study concluded that 80% of the countries in the world criminalize defamation and that defamation laws are being expanded to punish and threaten independent media. According to V-Dem, “Freedom of expression is the aspect of democracy undermined the most in autocratizing countries” (V-Dem Institute 2022). However, free speech is also under threat in open democracies. According to V-Dem, free expression saw declines in both Western Europe and North America. To mention one example, UNESCO has documented that criminal defamation remains on the statutes of 20 out of 25 surveyed states in these regions.

Left unchecked, the deterioration of free speech threatens individual and group freedoms, civil society, and democratic institutions, as well as progress in science and philosophy. The free speech recession is often attributed to authoritarian populism and its crackdown on dissent, civil society participation,

and an independent press. While these are important contributing factors, the broader reasons might be complex. Governments and citizens of liberal democracies no longer seem to fully recognize the central value of free speech to the success of their societies and individual and community flourishing. Even in open societies, the democratization and virality of online speech are increasingly seen as a threat rather than a precondition for pluralist societies. Liberal democracies are experiencing a marked backsliding when it comes to free speech (online and offline) in the past decade (Kaye 2017; DW 2019; Freedom House 2022a; Mchangama 2022b; Reporters without Borders 2022). This has created a ripple effect by eroding international norms and inspiring and legitimizing draconian crackdowns in authoritarian states (Mchangama and Fiss 2019; Mchangama and Alkiviadou 2020a). For the moment, countries such as Germany and other European Union (EU) countries but also the EU itself are handing authoritarian regimes with laws such as the German Network Enforcement Act (NetzDG) and the proposed Digital Services Act (DSA; Act to Improve Enforcement of the Law in Social Networks 2017; Splittgerber and Walz 2021; European Parliament 2022). The 2017 NetzDG “alarmed human rights campaigners” (Erixon 2021) by imposing a legal obligation on social media companies to remove illegal content, including insult, incitement, and religious defamation, within a minimum of 24 hours and at risk of a fine of up to 50 million euros. Such laws are problematic in themselves and become even more so where a rule of law framework is lacking, authoritarianism marks the political landscape, and government censorship is common (Human Rights Watch 2018; Hassan 2020; McDonnell 2020; Baghdasaryan and Gullo 2021; Freedom House 2021b; Noyan 2021). Legislative measures such as the NetzDG contribute to a regulatory race to the bottom for free speech with social media platforms having become the ultimate arbiters of harm, truth, and the practical limits of free speech.

Private companies, not formally bound by international human rights law (IHRL) but instead driven by their own business models and affected by the enhanced state pressure to remove online speech, are now “content gatekeepers” (Sander 2020, 941). Today’s media landscape is altering predominantly due to the rise of digital media. A handful of social media platforms act as the gatekeepers through which people access news and information in an easy and attractive way. Schulz (2019) argued that the growing number of users relying on social media for news is positively correlated to their discontent with mainstream media. This “platformization” of global information streams is not without its troubles, because these platforms’ algorithms are tweaked to increase engagement and thus often increase the visibility of hate and disinformation. Within this premise, countries are placing increasing pressure on social media platforms to remove allegedly harmful content. The impact state pressure has had on these companies is demonstrated by the drastic increase in content removal over the last few years. Big platforms are spending “growing amounts of resources to police content and take down illegal, harmful and objectionable content” (Erixon 2021). Further, state pressure to remove even contentious areas of speech such as hate speech

quickly and at risk of fines has prompted companies to work proactively to preserve their business models (Dias Oliva 2020). However, algorithms may have biased data sets, placing already marginalized groups at risk of further silencing (Brown 2020). For example, an assessment of artificial intelligence (AI) tools for regulating harmful text found that African American English tweets are twice as likely to be labeled offensive compared to others (Dias Oliva 2020). Moreover, centralized censorship hinders autonomous and organic ways for different communities to counter (perceived) harmful speech. This paternalistic approach goes against the very essence of the exercise of freedom of expression, which includes the right to seek, receive, and impart information and ideas of all kinds. News from free and diverse media sources is a fundamental tenet of liberal democracy, a central aspect of the right to freedom of information and expression, and an enabler of public participation and dialogue.

As such, and emanating from the position that free speech plays a crucial role for democratic values, giving voice and agency to racial, religious, and sexual minorities, our chapter will provide an insight into the fundamental reasons why alternative methods to dealing with online hate speech (such as counter-narratives and counterspeech) are an absolute necessity if free speech is to be protected, shedding light on the serious impact which the current regulatory framework has on speech. Aware of the fact that the current regulatory trends cannot be dissolved from one day to the next, the chapter offers a section on how IHRL can be better integrated into the regulatory approach for purposes of ensuring a more speech-protective handling of online content.

The Global Decline of Free Speech and the Role of Social Media

Social media platforms have “created unprecedented possibilities for widespread cultural participation and interaction” (Sander 2020, 941). With 4.9 billion active social media users, this has been a great leap forward for humanity, empowering marginalized and traditionally silenced groups, enhancing connectivity, and allowing for awareness raising on, inter alia, human rights violations (Ruby 2023). In fact, in today’s digital reality, people have “little choice but to participate” (Sander 2020, 939) on online platforms. The use of social media is marked by phenomena such as hate speech, violent extremism, and disinformation, with tragic events including the abuse of platforms during the Rohingya genocide (Milmo 2021).

Hate speech has long been a topic of deliberation, legislation, and judicial decisions across continents and within IHRL, but its dissemination on private platforms has made the question of its definition and regulation even more acute. Similarly, especially after the 2016 US presidential election and the COVID-19 pandemic, concerns have been expressed about the viral spread of disinformation. These developments have also raised serious questions about the legitimacy of private platforms as the arbiters of truth and deciders of harm as well as practical concerns about the suitability of large-scale content

moderation and the dangers of privatized censorship. As a result, more and more countries have passed legislation that restricts online speech and requires social media platforms to remove illegal or “harmful” user content by means of intermediary liability (Mchangama, Alkiviadou, and Mendiratta 2021). The pressure from states, users, and civil society to police various forms of harmful content has also caused platforms to remove more and more content with every passing year, as evident by their increased takedown figures (discussed in the Introduction to this book). So, while platforms were once focused on expanding the limits of free speech and access to information, they now increasingly rein in the exercise of these freedoms.

Through content moderation policies, platforms define the limits of speech, an issue that has emerged as “one of the most pressing challenges for freedom of expression in the 21st century” (Sander 2020, 941). Private companies, driven by their own business models are now “content gatekeepers”, exerting immense influence over public discourse globally. However, the plot thickens since their role is enhanced by the escalating moderation responsibilities imposed by governments through legislation, making such platforms “even more powerful” (Erixon 2021).

Therefore, content moderation by social media companies is one of the key issues affecting the practical exercise of free expression around the world. The global nature of major social media platforms creates significant problems when it comes to determining where to draw the line on various categories of content. Through content moderation policies, platforms define the limits of speech, an issue that has emerged as one of the most pressing challenges for freedom of expression in the 21st century. In Europe, legitimate concerns about online hate speech, disinformation, and extremism have incentivized democratic governments to respond with illiberal and counterproductive regulations that undermine the very values they are supposed to protect. This even includes some Scandinavian countries that have traditionally sought to fight extremism primarily through democratic debate, education, and a robust and vigilant civil society.

Two reports released by Justitia demonstrate how the German NetzDG precedent has spilled over in more than 20 states beyond Europe, including authoritarian regimes such as Russia and Venezuela (Mchangama and Fiss 2019). Heightening pressure on private companies to remove content makes them “even more risk averse in their moderation policies”, thereby shrinking civic space and placing free speech in dire straits (Erixon 2021). The EU’s DSA also poses significant risks to free speech by enhancing platform liability and requiring companies to swiftly remove vague categories of contentious speech such as “illegal content”, which is likely to result in weaker protections of online speech than what follows under IHRL.

Big platforms already have terms that regulate permissible content beyond that which is illegal, with companies such as Facebook spending “growing amounts of resources to police content and take down illegal, harmful and objectionable content” (Erixon 2021). Justitia conducted a legal analysis of 63

million Facebook comments that found that of the 1.4% of the comments classified as “hateful attacks” by Facebook, only about 0.0066% actually violated provisions of the Danish Criminal Code (Mchangama and Callesen 2022). In another survey of the Facebook accounts of five Danish media outlets, we found that only 1.1% of deleted comments violated the criminal code (Mchangama, Vinther-Jensen, and Brandt Taarnbørd 2020). Justitia’s research shows that social media is hardly the “Wild West” (Sánchez Nicolás 2022) that large platforms are often perceived to be by prominent European politicians like Emmanuel Macron (Reuters 2021). Siegel et al. (2019) conducted a study to assess whether Trump’s 2016 election campaign (and the six-month period following it) led to a rise in hate speech on Twitter. Based on an analysis of a sample of 1.2 billion tweets, they found that between 0.001% and 0.003% of the tweets contained hate speech on any given day – “a tiny fraction of both political language and general content produced by American Twitter users” (Siegel et al. 2019, 86). Similarly, a recent study in Harvard’s *Misinformation Review* documented how, once President Trump’s election fraud Tweets were labeled as misinformation on Twitter, they gained more traction on other platforms (Sanderson et al. 2017).

It is important to acknowledge that some of the restrictions that form part of the free speech recession are well intentioned and result from genuine concerns about the harmful potential of disinformation and hatred sowing division and weaponizing social media against democratic institutions, tolerance, and equality. Free speech comes with harms and costs that can contribute to real life harms such as the attack on the Capitol on 6 January, the amplification of polarization, and the ability of extremist voices to coordinate. However, it does not follow that censorship is an appropriate or efficient remedy to combat such harms. Speech-restrictive policies adopted by governments and big tech companies and advocated by civil society organizations come with great risk of negative unintended consequences. Frequently the justification for restricting speech relies on anecdotal rather than empirical data and on untested assumptions rather than rigorous research (Deller 2017; Siegel et al. 2019). Moreover, too often speech-restrictive policies assume that these will cure harms without cost to democratic values, when there is evidence for the opposite (Mchangama 2022c). The European Commission’s push for new censorship powers cites dramatic increases in online anti-Semitism in France and Germany. Yet, the EU’s two most powerful countries already enforce some of the most speech-restrictive hate speech laws in any democracy. In 2023, *The New York Times* found more than 8500 German cases of hate speech investigations by the police, with more than 1000 people charged or punished since 2018 (Satarino and Schuetze 2022). This includes a climate activist whose house was raided and devices confiscated by the police after he accused a far-right politician of being Islamophobic on Twitter. The scope creep of such laws might also skew important societal debates in favor of particular viewpoints. In the UK, feminists and Christians have been charged with offending LGBTQIAP+ people for

arguing that there are biological differences between the sexes (Chapman 2022). In France, an LGBTQIAP+ rights organization was fined for calling an opponent of same-sex marriage a “homophobe” (Lang 2016).

Further, studies have shown that extremists who are deplatformed from mainstream social media for violating terms migrate elsewhere with fewer rules (Urman and Katz 2022). This may not only lead to further radicalization but also defeat law enforcement efforts and impede counternarrative efforts, which could plausibly be effective in reducing hate speech. Norwegian scholar Jacob Ravndal argued that the rise of far-right extremism in Western Europe emanates from a combination of high immigration, low electoral support for radical right political parties, and the “extensive public repression of radical right actors and opinions” (Ravndal 2018, 846). Although he noted that such repression may discourage people from joining extreme groups, it may also push others to follow more violent paths. Moreover, several studies suggest that improper and overbroad removals make some users suspicious and may counteractively reinforce falsehoods and violent extremism. A 2022 paper by Bartusevicius and van Leeuwen analyzed 101 nationally representative samples from three continents and revealed a positive association between perceived levels of repression and intentions to engage in anti-government violence. Additional analyses from three specific countries in the studies characterized by widespread repression and anti-government violence identified a strong positive association between personal experience with repression and intentions to engage in anti-government violence. These results suggest that political repression of speech, aside from being normatively abhorrent, creates psychological conditions for political violence. Similarly, another recent study published in Harvard’s *Misinformation Review* documented how, once President Trump’s election fraud Tweets were labeled as misinformation on Twitter, they gained more traction on other platforms (Sanderson et al. 2017). The study argued that Twitter’s labeling of certain Tweets was not only ineffective at preventing the spread of Trump’s claims but it might have even backfired at an ecosystem level by drawing additional attention to messages that Twitter deemed problematic. Nevertheless, the pressure on platforms to remove disinformation is steadily increasing. In research conducted by the Royal United Services Institute on the far-right group Britain First, scholars found that limiting the accessibility of extremists to Facebook reduced their interaction with others and the dissemination of their ideas. However, their migration to other platforms with less moderation led to their content becoming more extreme. As such, as highlighted by Swedish economist and writer Fredrik Erixon, there is a need to understand the “behavioural consequences that follow from heavy-handed approaches to content regulation” (Erixon 2021).

Laws enhancing (often quick) content removal by intermediaries, however well intended, often serve as a “Trojan Horse” for wider free speech restrictions (Mchangama, Alkiviadou, and Mendiratta 2021; Mchangama and Callesen 2022). These tend to undergo “scope creep” and sometimes even end up affecting the very values and principles they are supposed to protect, such as equality

and nondiscrimination. Instead, we demonstrate how the existence of free speech may contribute to a healthy society, with one of our studies even reflecting the link between free speech and less terrorism (Eskikdsen and Bjørnskov 2020).

Case Study: The EU's Digital Services Act

National and regional legislative measures/proposals that dramatically enhance platform liability for content developed by users such as NetzDG and the EU's proposed DSA place free speech at risk and potentially shrink civic space. Such measures render private companies, not bound by IHRL, arbiters of fact and law. To meet obligations and avoid hefty fines, social media platforms are adopting a "better safe than sorry approach", increasingly relying on AI to (proactively) remove even contentious areas of speech such as hate speech. Against the backdrop of the current developments in the form of the proposed DSA, this section provides an overview of the challenges that emanate from the current European approach with a particular emphasis on contested areas of speech such as hate speech and disinformation and puts forth proposals that can be taken into consideration during negotiations and discussions. While cognizant of the fact that the structural composition of the DSA, in particular its platform liability approach, will not change (for now), we put forth ideas that could feed into the negotiation process, namely, a rights-based approach to content moderation.

In January of this year, with 530 votes in favor, 78 against, and 80 abstentions, the European Parliament adopted the text of the DSA that will be used in negotiations with Member States. After five trilogues, on 22 April, the Council and Parliament reached a provisional political agreement on the DSA. The DSA entered into force on 16 November 2022. On 25 April 2023, the European Commission adopted the first designation decisions under the DSA, designating 17 very large online platforms (VLOPs) and two very large online search engines that reach at least 45 million monthly active users.

The impact of the DSA on individuals, groups, companies, states, civil society, and civic space in Europe and beyond cannot be stressed enough. These new sets of rules, which seek to circumvent the spread of illegal content online and enhance big tech transparency, are creating a "new digital world, shaping our society for decades to come" (Becker 2022). As Facebook whistle-blower Frances Haugen told the EU, the DSA could become a "global gold standard" in content moderation (Noyan 2021). In essence "the DSA will regulate how human rights are exercised or curtailed online" (Becker 2022). In fact, the DSA has the potential to become an "important tool in order to guarantee a proper protection of fundamental rights by sector specific legislation" (Barata 2021). It is imperative that the final negotiated text will stay true to the values enshrined in the Charter of Fundamental Rights, including the freedom of expression. The DSA is significant in terms of imposing transparency and accountability requirements on platforms and new user rights. In relation to transparency reports that will need to be issued by all intermediaries

(regardless of size), it is noted that several platforms already issue transparency/enforcement reports, but this content remains inadequate. Ranking Digital Rights' 2020 Corporate Accountability Index highlights that "the most striking takeaway is just how little companies across the board are willing to publicly disclose about how they shape and moderate digital content, enforce their rules, collect and use our data and deploy the underlying algorithms that shape our world" (Brouillette 2020). Further, under the DSA, all hosting providers must give reasonings for decisions on content moderation, establish an internal complaint handling mechanism, and partake in out-of-court dispute settlements. While the DSA introduces some new transparency rules that are "straightforward and desirable", such as transparency reports, other mechanisms are not as simple (Erixon 2021). For example, by obliging platforms to inform users whose content has been removed of the reasoning does protect freedom of expression, and on this level it is welcomed. However, considering the multitude of obligations under this regulation and the sheer amount of online content, it could be foreseen that platforms will prefer to remove content than maintain and provide reasoning. In relation to content removal, the text approved by the European Parliament establishes a "notice and action" process. Upon such notices, hosting services should act "without undue delay, taking into account the type of illegal content that is being notified and the urgency of taking action" (EDRi 2022). Rather than endorsing general monitoring obligations, Members of European Parliament voted in favor of maintaining conditional liability for online intermediaries, shielding them from liability for user-generated illegal content that is not brought to their attention (EDRi 2022). This is positive, and we urge negotiators to ensure that it stays this way. While it is indisputable that general monitoring obligations would have led to even more over-removals of legitimate speech than we are used to and despite conditional liability being a preferred mechanism, things are far from perfect. While the text adopted by the European Parliament does not directly impose further liability on platforms (by sticking to conditional liability), the very role endowed to private companies to make decisions on the fundamental right to free speech is problematic. Beyond that, it could be argued that enhanced liability is achieved through alternate means. Specifically, the additional due diligence rules for VLOPs in terms of annual risk assessments under the close eye of the Commission as well as the possibility of fines for noncompliance "is the same thing as diluting the liability exemption directly" (Erixon 2021). Further, Barata (2021) argued that the mere notification of alleged illegality should not create knowledge or awareness to kick start the notice and action process "unless the notified content reaches a certain threshold of obviousness of illegality". Platform assessments that consider key values of IHRL such as legality, proportionality, and necessity should be part and parcel of this process. Also, it does not seem necessary or proportionate (in terms of free speech) that all categories of content/speech "entail the same consequences". On a normative level, as argued by Keller (2022), the DSA's content moderation approach is based on "breaking down human behavior and its governance into rationalized, bureaucratized, calculable components". While this

is the approach adopted by large platforms, the DSA, in antithesis to platforms, seeks to add consistency and foreseeability to “evolving human behavior”. So are large platforms’ existing content moderation practices. In tandem with the notice and action process, the DSA stipulates that VLOPs should “assess the systemic risks” stemming from their functioning. This additional obligation will most probably have the same consequence as enhanced platform liability since the platforms may be prone to reducing such risks and subsequently reducing the possibility of a violation of the DSA (and the fines associated therewith; Erixon 2021). On a practical level, mitigating such risk will probably require the use of AI (with all the problems that come with this route as summed up above). Free expression, freedom from discrimination, and due process are all placed at risk when automated mechanisms come into play in the handling of contentious areas of speech. As such, mitigating systemic risks may also impact the exercise of freedom of expression, even within the framework of content that is illegal (albeit loosely defined by the DSA). The DSA recognizes four categories of systemic risks that should be assessed in-depth. The first deals with the amplification of illegal content (such as illegal hate speech) with our comments on this term put forth above. Another category that is of particular interest here concerns “any actual and foreseeable negative effects on the protection of public health ... or other serious negative effects to the person’s physical, mental, social and financial well-being”. Barata (2021) noted that the reference to negative effects is “not appropriate in terms of human rights law”. At the heart of the functioning of IHRL is the balancing of, at times, competing rights. Blanket bans on effects to generic areas such as financial well-being could not possibly meet any test of legitimacy, proportionality, or necessity. A major issue that must be highlighted is the broad definition of “illegal content” that is to be removed upon notification. The DSA holds that such content means “any information or activity ... which is not in compliance with Union law or the law of a Member State, irrespective of the precise subject matter or nature of that law”. The DSA also notes that the general idea that should underpin the concept of illegal content is that “what is illegal offline should also be illegal online” and that this should cover content including “hate speech” but also “unlawful discriminatory content.” Three themes are identifiable here. Firstly, that the DSA includes “a lot of constructive ambiguity” working with unclear definition, which “would require platforms to take a very cautious approach” (Erixon 2021). Barata (2021) argued that the DSA deliberately refrains from providing a sound definitional framework and that this “vagueness and broadness may trigger over-removals of content and affect the right to freedom of expression of users”. Secondly, that there is no accepted definition of “hate speech” among Member States. In fact, categories such as “hate speech” are given “divergent interpretations across the EU”. Thirdly, the fact that illegal content also extends to “unlawful” discriminatory content demonstrates the low threshold attached to what is to be deemed removable by intermediaries, an issue that contributes to the further jeopardization of freedom of expression. Further, mandating the removal of illegal content is achieved through a stringent monitoring system, particularly for “very

large” online platforms with more than 45 million users in the EU, who are at risk of penalties in cases of nonconformity with the DSA, contributing to the “better safe than sorry” approach discussed above, and contributing to the enhanced use of AI, with all of the challenges this carries. The DSA also provides for the appointment of a digital services coordinator in each Member State to ensure the application and enforcement of the DSA and who may investigate suspected infringements of certain duties by VLOPs. So, what happens in countries such as Hungary, which, in 2021, passed a Russia-inspired “gay propaganda” law banning the promotion of material on LGBTQIAP+ rights to minors in schools and the media (France 24 2021). President Duda of Poland described LGBTQIAP+ rights as an “ideology even more destructive” than the communist ideology that indoctrinated the Polish youth before 1989.² This country is following the Russian–Hungarian footsteps, with its lower house of its parliament having adopted a similar law. The danger of a catch-all provision prohibiting any illegal or discriminatory content is also reflected in the fact that several EU countries continue to maintain blasphemy and religious insult laws (although rarely implemented). For example, Article 525 of Spain’s Penal Code punishes, among others, the “vilification of religious feelings”. Or, for example, Article 283 of the Austrian Criminal Code, which punishes “publicly incite[ing] to commit a hostile act” against a church or religious community. So, how do we ensure that coordinators will uphold the values of the Charter of Fundamental Rights such as free speech and equality? To make matters worse, “trusted flaggers”, whose reports must be processed by intermediaries “expeditiously”, must be approved by digital services coordinators. Trusted flaggers must meet certain conditions, such as being objective and transparent in terms of funding. However, their very integration into the DSA, beyond the complexities of illiberal states, is cause for concern since this creates a two-path system of removal whereby private (nonjudicial non-state) companies (not bound by IHRL) are directed by nonjudicial and possibly non-state but potentially state-influenced entities to remove speech, with the former having to give priority to such requests. Further, the DSA states that due diligence obligations are adapted to the “type, nature and size” of the intermediary. The short- and long-term effects of this distinction can be a migration of content and users as described in the introductory section of this chapter. Users who are easily caught up by algorithms or humans for the undefined notion of hate speech may migrate to smaller platforms (as they are already doing) that are not under such stringent control (Urman and Katz 2022). Enhanced demands and liability imposed by the EU on VLOPs will “accelerate this development” (Erixon 2021). In light of the above problematic aspects of the DSA when it comes to freedom of expression in particular, as noted by Keller (2022), “lawmakers ... still have an opportunity to resist provisions that will be too little good to users”. As noted by McGowan (2021) when discussing the DSA, enhanced platform responsibility and liability “betrays the legislators’ goal of limiting corporate power over public discourse by formally assigning companies a role in deciding the legality of our speech”.

Intermediary Liability and Artificial Intelligence

As a response to enhanced regulatory requirements, due to the risk of steep fines, platforms are prone to taking the “better safe than sorry” approach and regulating content rigorously. However, as noted by Llansó (2020), online communication on such platforms occurs on a massive scale, rendering it impossible for human moderators to review all content before it is made available. The sheer quantity of online content also makes the job of reviewing even reported content a difficult task. To respond to both the need to dodge state fines and the technical aspect of content scale and quantity, social media platforms (SMPs) have increasingly relied on AI in the form of automated mechanisms that proactively or reactively tackle problematic content, including hate speech. In brief, as highlighted by Dias Oliva (2021, 701), AI provides SMPs with “tools to police an enormous and ever-increasing flow of information – which comes in handy in the implementation of content policies”. While this is necessary in areas involving, for example, child abuse and the non-consensual promotion of intimate acts among adults, the use of AI to regulate more contentious “grey” areas of speech, such as hate speech, is complex. In light of these developments, this chapter looks at the use of AI to regulate hate speech on SMPs, arguing that automated mechanisms, which may have biased data sets and be unable to pick up on the nuances of language, may lead to violations of the freedom of expression and the right to nondiscrimination of minority groups, thus further silencing already marginalized groups.

As noted by Dias Oliva (2020), relying on AI, even without human supervision, is a necessity when it comes to content that could never be ethically or legally justifiable, such as child abuse. However, the issue becomes complicated when it comes to contested areas of speech, such as hate speech, for which there is no universal ethical and legal positioning as to what it is and when (if at all) it should be removed. In the ambit of such speech, Llansó et al. (2020, 2) underlined that the use of AI raises “significant questions about the influence of AI on our information environment and, ultimately, on our rights to freedom of expression and access to information”. As Llansó et al. (2020, 8) pointed out, it poses “distinct challenges for freedom of expression and access to information online”. A Council of Europe report (2017) highlighted that the use of AI for hate speech regulation directly impacts freedom of expression, which raises concerns about the rule of law and, in particular, notions of legality, legitimacy, and proportionality. The Council of Europe noted that the enhanced use of AI for content moderation may result in overblocking and consequently place freedom of expression at risk (Helberger et al. 2019). Gorwa, Binns, and Katzenbach (2020) argued that the increased use of AI threatens to exacerbate already existing opacity of content moderation, further perplex the issue of justice online, and “re-obscure the fundamentally political nature of speech decisions being executed at scale”. Moreover, regardless of the technical specifications of a particular mechanism, proactive identification (and removal) of

hate speech constitutes prior restraint of speech, with all of the legal issues that this entails. Specifically, Llansó et al. (2020, 3) argued that there is a “strong presumption against the validity of prior censorship in international human rights law”. Former UN Special Rapporteur on the Freedom of Opinion and Expression (SRFOE) David Kaye expressed his concern about the use of automated tools in terms of potential over-blocking and argued that calls to expand upload filtering to terrorist-related and other areas of content “threaten to establish comprehensive and disproportionate regimes of pre-publication censorship” (Office of the High Commissioner for Human Rights [OHCHR] 2018a).

Dias Oliva (2020) argued that the use of AI may result in the biased enforcement of companies’ terms of service. This can be due to a lack of data and/or biased training data sets, leading to the potential silencing of members of minority communities (Llansó 2020). This can lead to violations of the freedom of expression and the right to nondiscrimination. In its report “Mixed Messages: The Limits of Automated Social Content Analysis”, the Centre for Democracy and Technology revealed that automated mechanisms may disproportionately impact the speech of marginalized groups (Duarte, Llansó, and Loup 2018). Although technologies such as natural language processing and sentiment analysis have been developed to detect harmful text without having to rely on specific words or phrases, research has shown that, as Dias Oliva (2021, 702) put it, they are “still far from being able to grasp context or to detect the intent or motivation of the speaker”. As noted by Dias Oliva (2020, 634), although hash-matching is widely used to identify child sexual abuse content, it is not easily transposed to other cases such as extremist content, which “typically requires assessment of context”. As Keller (2018, 7) explained, “No reputable experts suggest that filters are good enough to be put in charge of deciding what is illegal in the first place”.

In relation to this, Keller (2018, 3) noted that the decision of platforms to remove Islamic extremist content will “systematically and unfairly burden innocent internet users who happen to be speaking Arabic, discussing Middle Eastern politics or talking about Islam”. She referred to the removal of a prayer (in Arabic) posted on Facebook because it allegedly violated its Community Standards. The prayer read, “God, before the end of this holy day, forgive our sins, bless us and our loved ones in this life and the afterlife with your mercy almighty.”

Further, as found by Dias Oliva (2021, 702–703), such technologies are just not cut out to pick up on the language used by, for example, the LGBTQIAP+ community, whose “mock impoliteness” and use of terms such as “dyke”, “fag”, and “tranny” are a way of reclaiming power and a means for preparing members of this community to “cope with hostility”. They give several reports from LGBTQIAP+ activists on content removal, such as the banning of a trans woman from Facebook after she displayed a photograph of her new hairstyle and referred to herself as a “tranny”. Another example used by Dias Oliva (2020) is a research study that revealed that African American English tweets are twice as likely to be considered offensive compared to others, thus reflecting

the infiltration of racial biases in technology. Dias Oliva (2021, 705) pointed to the “confounding effects of dialect” that need to be taken into account to avoid racial biases in hate speech detection. This reflects the significance of contextualizing speech – something that does not bode well with the design and enforcement of automated mechanisms and that could pose risks to the online participation of minority groups.

Moreover, automated mechanisms fundamentally lack the ability to comprehend the nuance and context of language and human communication. For example, YouTube removed 6000 videos documenting the Syrian conflict (BBC News 2017). It shut down the Qasioun News Agency, an independent media group reporting on war crimes in Syria. Several videos were flagged as inappropriate by an automatic system designed to identify extremist content. As Dias Oliva (2020, 632) noted, other hash-matching technologies, such as PhotoDNA, also seem to operate in “context blindness”, which could be the reason for the removal of those videos. Facebook banned the word *kalar* in Myanmar, because radicals had given this word a “derogatory connotation” and used it to attack the Rohingya people in Myanmar. The word was picked up through automated mechanisms that deleted posts that may have used it in another context or with another meaning (including *kalar* oat, which means camel). This led to the removal of posts condemning the fundamentalist movements in the country.

International Human Rights Law

The recognition of IHRL, at least on a theoretical level, has been seen in approaches of major platforms such as Facebook. For example, in March 2021, Facebook launched its Corporate Human Rights Policy, which outlines the human rights standards as defined in international law and sets out how they will be applied to, among others, their policies (Meta 2021). The Oversight Board (a body of international experts created by Facebook in 2020 to make final decisions regarding content moderation questions, including the evaluation of complaints by users) has also embraced IHRL in judging the appropriateness of Facebook’s content moderation decisions. The Oversight Board, in deciding a variety of issues ranging from hate speech to nudity to dangerous individuals and organizations, has relied on relevant provisions and principles of IHRL. The IHRL benchmark is paramount for ensuring legitimacy to the restriction of freedom of expression, one that is emphatically missing from the DSA’s very structure and essence. The section on IHRL will thus follow suit on recommendations of the UN SRFOE and Justitia’s extensive report on IHRL as a framework of first reference. In addition to proposing ways in which IHRL can be better embedded in the EU’s vision for its digital future, we will discuss how decentralization of content moderation could boost end-user control over data and privacy and simultaneously protect freedom of expression. While we are wary that at the heart of the DSA lies enhanced intermediary liability, this

chapter also argues that at least some form and extent of a decentralized approach could be integrated in parallel.

As private entities, social media platforms are not signatories to or bound by such documents, but, as the former SRFOE David Kaye has argued, IHRL is a means to facilitate a more rights-compliant and transparent model of content moderation. At the same time, the global nature of IHRL may also prove useful in dealing with the differences in national perception and legislation that characterize the global ecosystem of online expression. Yet, applying IHRL to private companies is a difficult task involving a plethora of challenges and dilemmas. In this section we will argue that to be compliant with IHRL, a platform's content moderation practices must be legitimate, necessary, and proportional within the framework of Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR; restrictions on freedom of expression), which sets out the grounds for limitation of freedom of expression. For hate speech, platforms should frame terms and conditions based on a threshold established by and take strictly into consideration the Rabat Plan of Action's six-part threshold test for context, speaker, intent, content and form, extent of dissemination, and likelihood of imminent harm before taking any enforcement action. For disinformation, a platform's terms and conditions should be tailored to protect the grounds in Article 19(3) ICCPR and Article 25 ICCPR (right to participate in voting and elections). In addition, platforms must refrain from adopting vague blanket policies for removal. Only disinformation promoting real and immediate harm should be subject to the most intrusive restrictive measures such as content removal. In determining the limits of disinformation, platforms should focus on the post's content, its context, its impact, its likelihood of causing imminent harm, and the speaker's intent.

In his 2018 report on the regulation of user-generated online content, the former UN SRFOE David Kaye proposed a framework for content moderation that "puts human rights at the very centre" (OHCHR 2018b). He stressed that national laws are inappropriate given the geographical and cultural diversity of digital users and that IHRL provides a framework to address this central difficulty because it transcends national boundaries. Kaye stressed that relying on IHRL to determine acceptable and unacceptable speech "enables forceful normative responses against undue State restrictions – provided companies play by similar rules" (OHCHR 2018b). He pointed to the UN Guiding Principles on Business and Human Rights, which provide that private companies must respect IHRL, noting that this obligation "exists independently of States' abilities and/or willingness to fulfill their own human rights obligations and does not diminish those obligations (UN 2011). In her 2021 Report on Disinformation, David Kaye's successor as SRFOE, Irene Khan, called for multi-dimensional responses to disinformation that are grounded in the IHRL framework (OHCHR 2021). Several leading academics have echoed this sentiment. Evelyn Aswad (2018) argued that international law is the most suitable framework for protecting freedom of expression. Similarly, Susan Benesch

(2020, 90) suggested that, even though IHRL cannot be used as “right off the shelf”, it can be the framework for content moderation. Hilary Hurd (2019) underlined that, while Article 19 of the ICCPR only applies to states, there have been “renewed calls to apply Article 19 to technology companies”. A team of researchers drafted a set of 16 recommendations put forth by the Israel Democracy Institute and Yad Vashem, which provide policy guidelines and benchmarks for content moderation “anchored in the applicable human rights standards” (IDI-Yad Vashem 2020, 16). Barrie Sander (2020, 967) also endorsed the adoption of a human rights-based approach to content moderation since this would provide social media platforms with a “common conceptual language to identify the impact of their moderation rules”. Dias Oliva (2020, 617) built on this by arguing that the ICCPR provides a “methodology and vocabulary for platforms to analyze whether their content policies and decisions are reasonable”. In a report by Stanford’s Law and Policy Lab, Sarah Shirazyan and others provided extensive insight into international case law, national legislation, and social media content policies in relation to violent extremist organizations, misinformation and fake news, online defamation, and cyber harassment and bullying (Shirazyan et al. 2020). The Oversight Board (an independent private body created by Facebook in 2020 to make final decisions regarding content moderation questions, including the evaluation of complaints by users) has also embraced IHRL in judging the appropriateness of Facebook’s content moderation decisions (Oversight Board 2023). By adopting an IHRL approach to content moderation, private platforms would also accommodate user demand since many remain deeply skeptical about state regulation of social media. Justitia’s 2021 global survey on attitudes toward free speech showed that people in two-thirds of the 33 countries surveyed prefer the regulation of social media content to be carried out solely by the companies themselves, and a plurality in the rest prefer the regulation of content to be carried out by social media companies along with national governments (Skaaning and Krishnarajan 2021). That said, a number of scholars have also expressed important reservations about marrying IHRL and content moderation. Danielle Citron (2017) argued that IHRL is just too flexible to provide the level of clarity that social media platforms need. Although Evelyn Douek (2018) also underlined the problem of IHRL’s flexibility and notes that there is little that actually compels such platforms to adhere to IHRL, she argued that it has the potential to develop “more concrete rules”. Brenda Dvoskin (2020) took a more rigid approach, suggesting that adopting IHRL “might not lead to more legitimate content moderation rules” because IHRL is not neutral and “leaves many speech questions unanswered”. We recognize that using IHRL will not resolve all of the thorny issues and dilemmas related to content moderation by private platforms and that it is unrealistic to expect that all content moderation decisions will be compliant with IHRL, even if all major platforms were to adopt an IHRL approach. We also acknowledge that an IHRL approach to content moderation will result in a significantly more speech-protective social media

environment, leaving in place much content that is likely to be false and misleading and/or cause offense and be deemed unacceptable/hateful/harmful by various states and constituencies across the globe. We also recognize that an IHRL approach to content moderation on private platforms will necessarily have to be adapted to the specific circumstances of social media rather than copied wholesale for entities that are very different from states for which IHRL was developed to constrain and guide. Accordingly, an IHRL approach should be seen as an imperfect improvement rather than a perfect solution. Nevertheless, we believe that the adaption is possible and would be beneficial for moderating the hate speech and disinformation found on platforms in today's centralized social media environment. Moreover, accepting the current absence of a basic framework setting out global norms for free speech on platforms would be akin to granting absolute discretion to a number of private companies, which constitute the central agora of global and local expression and whose content moderation policies have an enormous impact on the practical limits of freedom of expression around the globe. This is particularly problematic in the many countries in which official censorship and propaganda leave social media as the only way to express and organize dissent.

Therefore, IHRL as a framework of first reference will thus provide platforms legitimacy in resisting rising demands by states – as well as non-state actors – to take down content that they claim is in violation of national laws but that may be protected under IHRL. Ultimately, the future of free speech online may be best served by a more decentralized media environment and/or through enhanced user control over content. However, until such decentralization is achieved, we believe that IHRL as a “framework of first reference” for major social media platforms may cultivate a more transparent, legitimate, and speech-protective approach to handling online hate speech and disinformation. Moreover, IHRL as a “framework of first reference” should be capable of coexisting with enhanced user control of content, which would offset the very real concerns that IHRL may allow content that some/many users find too offensive, hateful, or misleading to tolerate.

Conclusion

The free speech recession takes place amidst a perfect storm of rising authoritarianism and a revolution in communication technologies. This poses huge challenges for those who want free speech to thrive as a fundamental norm of global importance. These challenges include the practical limits of free speech increasingly being decided by content moderation of private tech companies, unbound by constitutional or human rights norms. Moreover, technologies such as machine learning offer automated censorship systems vastly more sophisticated and comprehensive than any previous censorship regime in human history. We are also witnessing a cross-fertilization of censorship between regime types where restrictive standards developed in democracies are being copy-pasted in authoritarian

regimes. These challenges require a renewed effort at explaining why and how free speech matters. Existing standards must be updated to reflect technological developments and defended to withstand authoritarianism.

Free speech is often accused of providing shelter to harmful content and fuel hate crimes, terrorism, religious extremism, and the spread of corrosive propaganda. Yet, very often these claims rely on anecdotal and “intuitive” evidence rather than robust, data-driven, and empirical research on the relationship between free speech and social outcomes. For instance, in Europe there has been a rush to ban Russian disinformation due to the invasion of Ukraine (Mchangama 2022a). But an army of Open Source Intelligence people use Russian state-sanctioned information to debunk Russian propaganda and to find nuggets of information that can be pieced together to provide a more accurate picture of what is happening on the ground. Banning such access to information may thus hurt those actors most likely to effectively combat Russian propaganda. Likewise, social entrepreneurs in countries like Venezuela are developing technological solutions that allow individuals and civil society to document, debunk, and resist disinformation in real time and thus fight back against official propaganda and censorship. When it comes to hate speech, counterspeech has proven promising, whereas censorious content moderation aimed at countering hate speech has often negatively affected dissidents relying on social media to counter official propaganda and censorship in countries like Syria and Afghanistan.

In sum, centralized censorship, an increasingly popular policy in the digital age, not only harms the speaker but also robs other users from accessing ideas/information/news that they may not otherwise encounter as well as imparting information on, for example, human rights violations that they are experiencing. Citizens of Global South countries are especially reliant on social media to receive and impart information beyond state censorship and control and, to this end, we are mindful of the spillover effect that European legislation on platform liability can have on authoritarian and semi-authoritarian states. To this end, and based on the above analysis, the chapter argues that the current status quo vis-à-vis the regulatory approach adopted by states and tech companies to handle phenomena such as hate speech online severely damages the right to freedom of expression but also impacts minority groups that such measures purportedly seek to protect.

A recipe of enhanced alternative methodology such as counterspeech and counternarratives, enhanced user control over content, and the integration of IHRL into the regulation of speech that meets ICCPR thresholds make up an effective package to reverse the current tide and narrative.

Notes

- 1 Note that this chapter is based on findings and reports of phase 1 of the Future of Free Speech Project available here: <https://futurefreespeech.com/publications/> (accessed 8 September 2023), and parts of this chapter are based on Alkiviadou (2022).
- 2 Andrzej Duda’s speech during his 2020 presidential campaign.

References

- Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act). 2017. 12 July. <https://perma.cc/7UCW-AA3A>. Accessed 1 April 2023.
- Alkiviadou, Natalie. 2022. “Artificial Intelligence and Online Hate Speech Moderation.” *International Journal of Human Rights* 19 (32). <https://sur.conectas.org/en/artificial-intelligence-and-online-hate-speech-moderation/>. Accessed 6 September 2023.
- Aswad, Evelyn Mary. 2018. “The Future of Freedom of Expression Online.” *Duke Law & Technology Review* 17 (1): 26–70. <https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1331&context=dltr>
- Baghdasaryan, Meri, and Karen Gullo. 2021. *UN Human Rights Committee Criticizes Germany’s NetzDG for Letting Social Media Platforms Police Online Speech*. Electronic Frontier Foundation. 23 November. <https://www.eff.org/deeplinks/2021/11/un-human-rights-committee-criticizes-germanys-netzdg-letting-social-media>. Accessed 13 April 2023.
- Barata, Joan. 2021. *The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations*. Plataforma por la Libertad de Informacion. <https://libertadinformacion.cc/wp-content/uploads/2021/06/DSA-AND-IT-S-IMPACT-ON-FREEDOM-OF-EXPRESSION-JOAN-BARATA-PDLL.pdf>. Accessed 1 May 2023.
- Bartusevicius, Henrikas, and Florian van Leeuwen. 2022. “Poor Prospects – Not Inequality – Motivate Political Violence.” *Journal of Conflict Resolution* 66 (7–8): 1393–1421. doi:10.1177/002200272211074647
- BBC News. 2017. “YouTube ‘Made Wrong Call’ on Syria Videos.” 23 August. <https://www.bbc.com/news/technology-41023234>. Accessed 21 April 2023.
- Becker, Sebastian. 2022. *Framing the Future of the Internet*. Social Europe. 20 January. <https://www.socialeurope.eu/framing-the-future-of-the-internet>. Accessed 1 May 2023.
- Benesch, Susan. 2020. “But Facebook’s Not a Country: How to Interpret Human Rights Law for Social Media Companies.” *Yale Journal on Regulation Online Bulletin* 38: 86–111. <https://ssrn.com/abstract=3692701>
- Brouillette, Amy. 2020. “Key Findings: Companies Are Improving in Principle but Failing in Practice.” *2020 Ranking Digital Rights Corporate Accountability Index*. <https://outlook.stpi.narl.org.tw/pdfview/4b1141007770c91c0177dd54d7ac79f6>. Accessed 1 May 2023.
- Brown, Annie. 2020. “Biased Algorithms Learn from Biased Data: 3 Kinds Biases Found in AI Datasets.” *Forbes*, 7 February. <https://www.forbes.com/sites/cognitiveworld/2020/02/07/biased-algorithms/>. Accessed 1 May 2023.
- Chapman, Ben. 2022. “Media Commentator Caroline Farrow Arrested by Police over Gender Twitter Spat.” 4 October. <https://www.gbnews.com/news/media-commentator-caroline-farrow-arrested-by-police-over-gender-twitter-spat/373269>. Accessed 1 May 2023.
- Citron, Danielle Keats. 2017. “What to Do about the Emerging Threat of Censorship Creep on the Internet.” 28 November. <https://www.cato.org/policy-analysis/what-do-about-emerging-threat-censorship-creep-internet>. Accessed 1 May 2023.
- Committee to Protect Journalists. 2023. “Deadly Year for Journalists as Killings Rose Sharply in 2022.” <https://cpj.org/reports/2023/01/deadly-year-for-journalists-as-killings-rose-sharply-in-2022/>. Accessed 10 May 2023.
- Council of Europe. 2017. “Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications.” Council of Europe Study DGI(2017)12. <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>. Accessed 29 April 2023.

- Deller, Rose. 2017. "Book Review: Hate Speech and Democratic Citizenship by Eric Heinze." *LSE Review of Books*, 23 March. <https://blogs.lse.ac.uk/lserewiewofbooks/2017/03/23/book-review-hate-speech-and-democratic-citizenship-by-eric-heinze/>. Accessed 23 April 2023.
- Dias Oliva, Thiago. 2020. "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression." *Human Rights Law Review* 20 (4): 607–640. doi:10.1093/hrlr/ngaa032
- Dias Oliva, Thiago. 2021. "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online." *Sexuality & Culture* 25: 700–732. doi:10.1007/s12119-020-09790-w
- Douek, Evelyn. 2018. "U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for Human Rights by Default." *Lawfare*, 6 June. [https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default#:~:text=The%20report%20calls%20for%20states,for%20people%20harmed%20by%20moderation](https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default#:~:text=The%20report%20calls%20for%20states,for%20people%20harmed%20by%20moderation.). Accessed 1 May 2023.
- Duarte, Natasha, Emma J Llansó, and Anna Loup. 2018. "Mixed Messages? The Limits of Automated Social Media Content Analysis." Paper presented at the 1st Conference on Fairness, Accountability, and Transparency, New York. <https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf>. Accessed 4 May 2023.
- Dvoskin, Brenda. 2020. *International Human Rights Law Is Not Enough to Fix Content Moderation's Legitimacy Crisis*. Berkman Klein Center Collection. 16 September. <https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd>. Accessed 1 May 2023.
- DW. 2019. "Dozens of Raids in Germany over Online Hate Speech." 6 June. <https://www.dw.com/en/germany-dozens-of-raids-over-online-hate-speech/a-49080109>. Accessed 15 April 2023.
- EDRi. 2022. "European Parliament Approves Rights-Respecting DSA & Proposes Ban on Use of Sensitive Personal Data for Online Ads." 20 January. <https://edri.org/our-work/european-parliament-approves-rights-respecting-dsa-proposes-ban-on-use-of-sensitive-personal-data-for-online-ads/>. Accessed 1 May 2023.
- Erixon, Fredrik. 2021. *Too Big to Care or Too Big to Share: The Digital Services Act and the Consequence of Reforming Intermediary Liability Rules*. <https://ecipe.org/publications/digital-services-act-reforming-intermediary-liability-rules/>. Accessed 19 April 2023.
- Eskikdsen, Lasse, and Christian Bjørnskov. 2020. "Does Freedom of Expression Cause Less Terrorism?" *Political Studies* 70 (1): 131–152. doi:10.1177/0032321720950223
- European Parliament. 2022. *Amendments Adopted by the European Parliament on 20 January 2022 on the Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC (COM(2020)0825 – C9–0418/2020 – 2020/0361(COD))*. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.html. Accessed 3 April 2023.
- France 24. 2021. "Hungary's Controversial Anti-LGBT Law Goes into Effect Despite EU Warnings." 7 July. <https://www.france24.com/en/europe/20210707-hungary-s-controversial-anti-lgbt-law-goes-into-effect-despite-eu-warnings>. Accessed 1 May 2023.
- Freedom House. 2021a. *Freedom on the Net 2021*. <https://freedomhouse.org/country/belarus/freedom-net/2021>. Accessed 11 April 2023.
- Freedom House. 2021b. *The Global Drive to Control Big Tech*. <https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech>. Accessed 1 May 2023.
- Freedom House. 2022a. *Freedom in the World 2022*. https://freedomhouse.org/sites/default/files/2022-02/FIW_2022_PDF_Booklet_Digital_Final_Web.pdf. Accessed 14 April 2023.

- Freedom House. 2022b. *Freedom on the Net 2022*. <https://freedomhouse.org/sites/default/files/2022-10/FOTN2022Digital.pdf>. Accessed 12 May 2023.
- Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1). doi:10.1177/2053951719897945
- Hassan, Syed Raza. 2020. "Pakistani Journalist Arrested for Defaming Military." Reuters. 12 September. <https://www.reuters.com/article/pakistan-journalist-arrest-iDINKBN263058>. Accessed 30 April 2023.
- Helberger, Natali, Sarah Eskens, Max van Drunen, Mariella Bastian, and Judith Moeller. 2019. *Implications of AI-Driven Tools in the Media for Freedom of Expression*. Institute for Information Law (IViR). <https://rm.coe.int/coe-ai-report-final/168094ce8f>. Accessed 1 May 2023.
- Hurd, Hilary. 2019. "How Facebook Can Use International Law in Content Moderation." *Lawfare*, 30 October. <https://www.lawfareblog.com/how-facebook-can-use-in-international-law-content-moderation>. Accessed 1 May 2023.
- Human Rights Watch. 2018. *No Support: Russia's Gay Propaganda Law Imperils LGBT Youth*. <https://www.hrw.org/report/2018/12/12/no-support/russias-gay-propaganda-law-imperils-lgbt-youth>. Accessed 18 April 2023.
- IDI-Yad Vashem. 2020. "A Proposed Basis for Policy Guidelines for Social Media Companies and Other Internet Intermediaries." In *Reducing Online Hate Speech: Recommendations for Social Media Companies and Internet Intermediaries*, edited by Yad Vashem, 13–24. Jerusalem: Center for Democratic Values and Institutions. <https://en.idi.org.il/publications/31767>
- Kaye, David. 2017. "How Europe's New Internet Laws Threaten Freedom of Expression." 18 December. <https://www.foreignaffairs.com/articles/europe/2017-12-18/how-europes-new-internet-laws-threaten-freedom-expression>. Accessed 12 April 2023.
- Keller, Daphne. 2018. "Internet Platforms: Observations on Speech, Danger and Money." Hoover Institution's Aegis Paper Series, No. 1807. <https://ssrn.com/abstract=3262936>
- Keller, Daphne. 2022. "The DSA's Industrial Model for Content Moderation." *Verfassungsblog*, 24 February. <https://verfassungsblog.de/dsa-industrial-model/>. Accessed 1 May 2023.
- Lang, Nico. 2016. "French Hate Crime Ruling Sets a Dangerous Precedent for LGBT People: It's Now Illegal to Call Someone a 'Homophobe' in France." 7 November. <https://www.salon.com/2016/11/07/french-hate-crime-ruling-sets-a-dangerous-precedent-for-lgbt-people-it-is-now-illegal-to-call-someone-a-homophobe-in-france/>. Accessed 1 May 2023.
- Llansó, Emma. 2020. "No Amount of AI in Content Moderation Will Solve Filtering's Prior-Restraint Problem." *Big Data & Society* 7 (1). doi:10.1177/2053951720920686
- Llansó, Emma, Joris van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. "Artificial Intelligence, Content Moderation and Freedom of Expression." 26 February. <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf>. Accessed 26 April 2023.
- McDonnell, Stephen. 2020. "Ren Zhiqiang: Outspoken Ex-Real Estate Tycoon Gets 18 Years Jail." BBC News. 22 September. <https://www.bbc.com/news/world-asia-china-54245327>. Accessed 30 April 2023.
- Mchangama, Jacob. 2022a. "The Problem with Banning Russian Disinformation." *The Daily Beast*, 14 March. <https://www.thedailybeast.com/the-problem-with-banning-russian-disinformation>. Accessed 4 April 2023.
- Mchangama, Jacob. 2022b. "The War on Free Speech – Censorship's Global Rise." 9 February. <https://www.foreignaffairs.com/articles/world/2022-02-09/war-free-speech-censorship>. Accessed 11 May 2023.

- Mchangama, Jacob. 2022c. "Will Banning Hate Speech Make Europe Safer?" *The Wall Street Journal*, 4 February. <https://www.wsj.com/articles/will-banning-hate-speech-make-europe-safer-11643985138>. Accessed 1 May 2023.
- Mchangama, Jacob, and Natalie Alkiviadou. 2020a. *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two*. Justitia. https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf. Accessed 2 April 2023.
- Mchangama, Jacob, and Natalie Alkiviadou. 2020b. "In Search for an Antidote: COVID-19 Impact on Freedom of Expression" *Verfassungsblog*, 9 July. <https://verfassungsblog.de/in-search-for-an-antidote/>. Accessed 1 April 2023.
- Mchangama, Jacob, Natalie Alkiviadou, and Raghav Mendiratta. 2021. *A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of "Platformization."* Justitia. https://futurefreespeech.com/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf. Accessed 20 April 2023.
- Mchangama, Jacob, and Lucas Callesen. 2022. *The Wild West? Illegal Comments on Facebook*. Justitia. 20 January. <https://justitia-int.org/en/the-wild-west/>. Accessed 25 April 2023.
- Mchangama, Jacob, and Joelle Fiss. 2019. *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship*. Justitia. http://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf. Accessed 2 April 2023.
- Mchangama, Jacob, Eske Vinther-Jensen, and Ronne Brandt Taarnbørd. 2020. *Digital Freedom of Speech and Social Media*. Justitia. 29 May. <http://justitia-int.org/en/new-report-digital-freedom-of-speech-and-social-media/>. Accessed 21 May 2023.
- McGowan, Iverna. 2021. "European Plans to Regulate Internet Will Have Major Impacts on Civic Space at Home and Abroad." <https://cdt.org/insights/european-plans-to-regulate-internet-will-have-major-impacts-on-civic-space-at-home-and-abroad/>
- Meta. 2021. *Corporate Human Rights Policy*. <https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf>. Accessed 1 May 2023.
- Milmo, Dan. 2021. "Rohingya Sue Facebook for £150bn over Myanmar Genocide." *The Guardian*, 6 December. <https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence>. Accessed 2 May 2023.
- Noyan, Oliver. 2021. "Content Moderation Policies Continue to Face Core Dilemmas." 17 November. <https://www.euractiv.com/section/politics/news/content-moderation-policies-continue-to-face-core-dilemmas/>. Accessed 10 April 2023.
- Office of the High Commissioner for Human Rights. 2018a. *Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Expression*. Geneva: UN Human Rights Council.
- Office of the High Commissioner for Human Rights. 2018b. *A/HRC/38/35: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Geneva: UN Human Rights Council.
- Office of the High Commissioner for Human Rights. 2021. *A/HRC/47/25: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Geneva: UN Human Rights Council.
- Oversight Board. 2023. *Case Decisions and Policy Advisory Opinions*. <https://www.oversightboard.com/decision/>. Accessed 1 May 2023.
- Ravndal, Jacob Aasland. 2018. "Explaining Right-Wing Terrorism and Violence in Western Europe: Grievances, Opportunities and Polarisation." *European Journal of Political Research* 57: 845–866. doi:10.1111/1475-6765.12254

- Reporters without Borders. 2022. *Europe – Central Asia: Press Freedom in Europe Overshadowed by the War in Ukraine*. <https://rsf.org/en/europe-central-asia>. Accessed 14 April 2023.
- Reuters. 2021. “Macron: EU Will Draw Up Regulation to Fight Hate on Social Media.” 9 December. <https://www.reuters.com/business/macron-eu-will-draw-up-regulation-fight-hate-social-media-2021-12-09/>. Accessed 4 May 2023.
- Ruby, Daniel. 2023. *Social Media Users in the World (2023 Demographics)*. <https://www.demandsage.com/social-media-users/#:~:text=So%20to%20answer%20your%20question,are%20on%20social%20media%20platforms>. Accessed 1 May 2023.
- Sánchez Nicolás, Elena. 2022. “MEPs to Crackdown on Digital Wild West.” *EU Observer*, 20 January. <https://euobserver.com/democracy/154141>. Accessed 15 May 2023.
- Sander, Barrie. 2020. “Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation.” *Fordham International Law Journal* 43 (4): 939–1006. <https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3>
- Sanderson, Zeve, Megan A Brown, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2017. “Twitter Flagged Donald Trump’s Tweets with Election Misinformation: They Continued to Spread Both on and off the Platform.” *Harvard Kennedy School Misinformation Review*, 24 August. <https://misinforeview.hks.harvard.edu/article/twitter-flagged-donald-trumps-tweets-with-election-misinformation-they-continued-to-spread-both-on-and-off-the-platform/>. Accessed 20 April 2023.
- Satarino, Adam, and Christopher F Schuetze. 2022. “Where Online Hate Speech Can Bring the Police to Your Door.” *New York Times*, 23 September. <https://www.nytimes.com/2022/09/23/technology/germany-internet-speech-arrest.html>. Accessed 1 May 2023.
- Schulz, Anne. 2019. “Where Populist Citizens Get the News: An Investigation of News Audience Polarization along Populist Attitudes in 11 Countries.” *Communication Monographs* 86 (1): 88–111. doi:10.1080/03637751.2018.1508876
- Shirazyan, Sarah, Allen Weiner, Yvonne Lee, Madeline Magnuson, Amélie-Sophie Vavrovsky, Anirudh Jain, Asaf Zilberfarb, David Jaffe, Eric Frankel, Jasmine Shao, et al. 2020. *How to Reconcile International Human Rights Law and Criminalization of Online Speech: Violent Extremism, Misinformation, Defamation, and Cyberharassment*. Stanford Law School Law and Policy Lab. <https://law.stanford.edu/publications/how-to-reconcile-international-human-rights-law-and-criminalization-of-online-speech-violent-extremism-misinformation-defamation-and-cyberharassment/>
- Siegel, Alexandra, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2019. “Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and Its Aftermath.” *Quarterly Journal of Political Science* 16 (1): 71–104. doi:10.1561/100.00019045
- Skaaning, Svend-Erik, and Suthan Krishnarajan. 2021. *Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech*. Justitia. May 2021. https://futurefreespeech.com/wp-content/uploads/2021/06/Report_Who-cares-about-free-speech_21052021.pdf. Accessed 1 May 2023.
- Splittergerber, Andreas, and Caroline Walz. 2021. “EU: New Hate Speech Rules for Social Networks in the European Union.” ReedSmith. 29 June. <https://www.technologylawdispatch.com/2021/06/privacy-data-protection/eu-new-hate-speech-rules-for-social-networks-in-the-european-union/>. Accessed 5 April 2023.
- UNESCO. 2022. *Defamation Laws and SLAPPs Increasingly “Misused to Curtail Freedom of Expression.”* 8 December. <https://www.unesco.org/en/articles/defamation-laws-and-slapps-increasingly-misused-to-curtail-freedom-expression>. Accessed 13 April 2023.
- United Nations. 2011. *Guiding Principles on Business and Human Rights*. https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf. Accessed 4 April 2023.

- Urman, Aleksandra, and Stefan Katz. 2022. "What They Do in the Shadows: Examining the Far-Right Networks on Telegram." *Information, Communication & Society* 25 (7): 904–923. doi:10.1080/1369118X.2020.1803946
- V-Dem Institute. 2021. *Autocratization Turns Viral (Democracy Report 2021)*. https://www.v-dem.net/static/website/files/dr/dr_2021.pdf. Accessed 1 May 2023.
- V-Dem Institute. 2022. *Democracy Report 2022*. https://v-dem.net/media/publications/dr_2022.pdf. Accessed 12 May 2023.

6

ONLINE HATE SPEECH IN VIDEO GAMES COMMUNITIES

A Counter Project

Susana Costa, Bruno Mendes da Silva and Mirian Tavares

Introduction

Video games represent one of the most influential media in popular culture. Figures speak for themselves: on the European console market alone, the 20 best-selling games have generated more than 973 million copies (GamesIndustry.biz 2020). Each year, the size of the gaming communities has reached new records: according to the SuperData annual report (SuperData 2021), every year the video gaming industry growth is 12%. In 2021, there were more than 2.5 billion players in Europe, 70% aged under 18 (Council of Europe 2021). In 2020, the revenue from the video game industry worldwide was around €145 billion (Council of Europe 2021). Moreover, increasing numbers of players gather in virtual communities where they keep up to date with the latest news about their favorite games, follow one or more players, share game strategies, form teams, follow live streams, or share their own content. These virtual communities often create a sense of belonging, with certain norms and values (Tardini and Cantoni 2005; Piertersen et al. 2018; Rivera-Vargas and Mino-Puigercos 2018), where the relationship between the player, the community, and the game can create different identities (Gee 2003; Tardini and Cantoni 2005). In 2020, the monthly average of minutes watched on *Twitch*, one of the most popular streaming platforms dedicated to video games, including broadcasts of video game competitions, was 93 billion. This corresponds to almost 176,820 years. Every day, more than 26 million visitors watch *Twitch* content (Council of Europe 2021).

Previous findings identify the positive motivations for the game: socialization, fantasy, the opportunity to exercise control, fun, meaningful opportunities for social interaction, inspiring creativity, meaningful and lasting relationships, discovering a sense of community, challenge, and relaxation (Yee 2007; Dauriat

et al. 2011; Shi et al. 2019; Arbeau et al. 2020). In addition to identifying the reasons that motivate players, video games can prove to be excellent educational tools, able to motivate children and youth to learn a certain skill based on the development of critical thinking, cooperation, interaction, and engagement (Scolari 2018). Researchers like Felicia (2009), Gee (2003), Bogost (2011), Frasca (1999), or Jenkins et al. (2006) advocate participatory, interactive, and digital learning frameworks, suited to the skills and interests of new generations. The games' ambiance is the favored environment for entertainment, socializing, communicating, and learning, reflecting the actual importance and influence of this specific media.

Yet, a decade of research studies on online hate speech shows that games and gaming communities are breeding ground for harmful content, racist expression, out-group hatred, online propaganda, sexism, and sexual discrimination (Breuer 2017; Bliuc et al. 2018; Blaya 2019; Gámez-Guadix, Wachs, and Wright 2020; Harriman et al. 2020). It urges analysis of the phenomenon of online hate speech, because studies show that exposure to and victimization by this toxic language have a negative impact on young players and are correlated with negative emotions, depression, anxiety, deviant behavior, and decreased well-being (Wachs and Wright 2018, 2019; Gámez-Guadix, Wachs, and Wright 2020; Gómez-García, Paz-Rebollo, and Cabeza-San-Deogracias 2021).

Despite the importance of the videogame industry and its role in young people's lives, the existing tools still struggle to fully tackle the problem of hate speech. In this study, we intend to (1) analyze how youth perceive online hate speech in video games and gaming communities and (2) look for strategies to prevent this phenomenon in the games' environment. The findings in this study aim to pave the way for a program to counteract online hate speech in video games.

Online Hate Speech

Online hate speech, also known as online toxic behavior, grieving, or online disinhibition (Suler 2004; Kwak and Blackburn 2014), is defined by European Union law "as the public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, color, religion, descent and national or ethnic origin" (Eur-lex 2008). In the last decade, the academic interest in hate speech saw an important increase, reflected, for example, in the volume of Web of Science-indexed production, which increased from 42 to 162 between 2013 and 2018 (Paz, Montero-Díaz, and Moreno-Delgado 2020). Researchers on the topic describe it as a set of behaviors that one categorizes as toxic in relation to constantly renegotiated and evolving social norms (Blaya 2019; Deslauriers, St-Martin, and Bonenfant 2020).

It is possible to identify different types of online hate speech motivated by gender, sexual identity, nationality, historical events, or religious beliefs (Blaya 2019; Paz, Montero-Díaz, and Moreno-Delgado 2020). The HateBase platform (HateBase 2023), a web-based application that collects instances of hate speech

online worldwide, indicates that the majority of cases of hate speech target individuals based on ethnicity and nationality, but incitements to hatred focusing on religion and class have also been on the rise.

To prevent the spread of illegal hate speech, on 31 May 2016, the European Commission together with Facebook, YouTube, Twitter, and Microsoft, agreed on a Code of Conduct on countering illegal hate speech online “in the effort to respond to the challenge of ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally” (European Commission 2016). The last evaluation of this Code of Conduct, published in June 2021 and carried out by the Directorate-General for Justice and Consumers, concluded that sexual orientation is the most reported ground of hate speech (33.1%), followed by xenophobia (including anti-migrant hatred; 15%) and anti-gypsyism (9.9%).

Experiences with hate speech online happen through three routes: exposure, victimization, and aggression (MacHackova et al. 2020). A report focusing on children from 11 to 17 years old and their experiences with cyberhate in 10 European countries concludes that exposure to hate speech increases with age, a tendency probably correlated to the overall higher engagement in online usage (MacHackova et al. 2020).

Online Hate Speech in Video Games

Multiplayer games provide players with the thrill of true competition (Kwak and Blackburn 2014), leading to the verbal expression of blasphemies and obscenities, often tolerated as a common reaction in moments of anger and frustration (Citron and Norton 2011; Gagliardone et al. 2015; Breuer 2017). During games, interaction in chats is common, and the dialogues diverge between praise and negative or ironic comments about performance in the game and personal insults based on sexual orientation or ethnicity, situations of harassment and attacks on minorities (Uyheng and Carley 2021). Hate speech in digital games is often the result of these interactive dynamics among players, in unmoderated activities, such as team building, sharing strategies, and chats or live streaming on game platforms and game communities, which are a common medium to spread this type of toxic behavior (Matamoros-Fernández 2017).

Dissociative anonymity, invisibility, asynchronicity, dissociative imagination, minimization of status and authority, individual differences, and predisposition may trigger toxic disinhibition (Suler 2004), which can favor the use of hate speech, characterized by the demonstration of power or expression of frustration in the face of defeat. This type of behavior can prove to be detrimental to the physical and psychological well-being and the self-esteem of aggressors and victims (Breuer 2017; Arbeau et al. 2020; Harriman et al. 2020).

The desensitization process, as a long-term consequence, has been studied since the 1960s, bringing up that exposure to violence and hatred in the media decreases the intensity of the emotional response of children and youth. Over

the years, different experiments have analyzed physical and psychological responses to hate, like heart rate and perspiration, along with emotional reactions and social cognitions, allowing the construction of a theoretical framework of desensitization with effects on personality and on how children and youth cope with hatred (Funk et al. 2004; Brockmyer 2013). More recent experiments showed the process of desensitization to online hate speech (Soral, Bilewicz, and Winiewski 2018; Uyheng and Carley 2021), confirming that the increased exposure to cyberhate leads to desensitization to this type of phenomenon, further showing that there is greater indifference toward the victims and the normalization of prejudice.

It is understood that the phenomenon of online hate speech is transversal to all online media; however, it seems important to understand how this type of speech is being manifested in the specific and growing field of digital games, nowadays part of most children's lives, by analyzing how children and youth perceive it and how this experience affects their behavior inside and outside the digital life.

Counternarratives to Counteract Online Hate Speech

Narratives have occupied a central place in the history of human culture since its beginnings. They are also one of the primary cognitive mechanisms for understanding the world and one of the fundamental ways in which communities are built (Murray 2003). Finally, we underline narratives as a powerful educational tool, for both teaching and learning.

Digital narratives, which are told in participatory systems, using computational technologies, involve the users in a different approach. According to Murray (2003), digital narratives make it possible to divide the experience into three levels: (1) immersion (journey to another reality), (2) agency (action within the narrative), and (3) transformation (effect of the narrative journey on the user). Through the digital narratives, players are transported to the virtual world and are able to interact with, participate in, and transform the space.

As the narrative becomes progressively richer and more complex, the greater the possibility for audience interaction and participation and the creation of one's own narrative. Considering the connectivity and participation characteristics of contemporary society, contents are often consumed and produced in a collaborative environment within a community, without physical barriers.

The process of creating a narrative is representative of the convergent and participatory culture (Jenkins et al. 2006), and it reflects the transformations that occur in the way of producing and consuming narratives, through the convergence of the media and the technological development of the platforms, boosted by digitization and interchangeability of media.

Narratives can be used to create alternatives to dominant narratives (alternative narratives) or to counter narratives (counternarratives). Both aim to deconstruct and discredit oppressive narratives (extremist, racist, xenophobic, or any other type of propaganda that affects individual freedom), promoting

values such as respect, openness, equality, understanding, and solidarity. It is a way to deconstruct or delegitimize a certain type of discourse that undermines the dignity of the other. Counternarratives have been developed and delivered by a variety of state and non-state actors (Blaya 2019).

Tuck and Silverman (2016) affirmed that to create effective counternarratives, it is necessary to consider factors such as age and language, offering content capable of generating thoughts, feelings, memories, and reflections. The authors argued that the creation of counternarrative content requires the expansion, redirection, and re-creation of existing content.

So far, a limited number of studies have been conducted on counternarratives as a response to the massive growth of online hate speech. Studies have focused on identifying successful counternarratives (Chung et al. 2021), evaluating their effectiveness, and the characteristics of counternarratives (Mathew et al. 2019).

Tekiroğlu, Chung, and Guerini (2020) distinguished eight types of reaction groups of counternarratives to combat online hate speech: (1) the presentation of facts against hate speech; (2) the presentation of contradictions in hate speech; (3) the notice of the offline or online consequences of hate speech; (4) affiliation with a given characteristic of the speaker, seeking empathy and dissuasion; (5) in denouncing hateful speech, through the mechanisms of digital platforms; (6) humor; (7) a positive tone; and (8) hostility.

Despite the importance of this analysis for understanding and preventing the phenomenon of online hate speech, it is considered that to deepen the study and manifestation of counternarratives it is necessary to establish a data set and develop a model. This approach was made in 2019, by the CONAN project (COunter NARRatives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech).

CONAN identified a gap in the availability of large-scale, appropriate counternarrative data sets, making it difficult to address the problem, and sought to establish the first large-scale, multilingual (English, French, and Italian) hate speech/counternarrative data set. The data set consisted of 4078 pairs in the three languages. Along with the hate speech/counternarrative pairs, a set of metadata was made available, such as demographics, hate speech subtheme, and type of counternarrative.

Most approaches to hate speech manifest themselves through silencing. The CONAN project sought to provide study data to an artificial intelligence system and validate its responses during the trial-and-error process. The human value of this system enabled the addition of specialized and sensitive knowledge, combined with computing speed. This collaborative dynamic helped reduce the weaknesses of both individuals and machines to achieve more accurate results.

The analysis of these studies and projects, based on the use of counternarratives to respond to hate speech (Citron and Norton 2011), allows us to conclude that by implementing a methodology capable of incorporating partnerships, content creation, and implementation of strategies based on counternarratives with the public, it seems possible to achieve a measurable and replicable impact.

Material and Methods

This study was carried out in the scope of the European project *Play Your Role: Gamification against Hate Speech*. The starting point of the project, a multilingual initiative, comprising five European partners, was the study of the state of the art and the collection of data. In a second phase, a set of tools was designed and created to combat online hate speech. The name of the project establishes a semantic relationship with the concept of role-playing – becoming someone else, somewhere else – and with the idea that events are triggered by choices made by the player – one of the main goals of the project.

To carry out this study, a questionnaire survey with closed questions was prepared, with answers scored on a Likert scale, where respondents were asked to specify their level of agreement, frequency, importance, and likelihood on a symmetric scale for a series of statements related to gaming and hate speech. The questionnaires were created in English and were then translated into the national languages of three countries: Portuguese, Lithuanian, and Italian. These surveys were approved by the educational supervision structures in each country, and they were applied by direct administration in six schools between September 2020 and March 2021. Due to the pandemic and school closures, the time to complete the surveys was more extended than anticipated. The questionnaire was implemented on paper during school hours. The topic was introduced to respondents by the definition of online hate speech proposed by the European Union and an audiovisual resource exposing some fact-checking about toxicity in digital gaming ambiances. Prior to implementation, the survey was tested by a pilot group of students and revised according to some suggestions.

This anonymous questionnaire was composed of five groups: in the first group, the goal was to understand the relationship between adolescents and the gaming experience (where they play, how long, with whom); the second group aimed at interpreting young people's perceptions and experiences of hate speech in online gaming communities; in group three, the questions concerned the use of live streaming and chat platforms for gamers and the contact with cyberhate in those virtual spaces; group four focused on the perception of the responsibility for the growth of this specific trend of online hate speech; and, finally, group five gathered new insight into how young people copy hate speech in video games and online communities.

The answers to the survey were analyzed using SPSS, and descriptive statistics and correlations between the variables were carried out, with no omission cases registered.

Hypothesis

Time of Play and Exposure to Hate Speech

Previous studies correlated the time youth spent online with the increasing probability of exposure to hate in the online space (Oksanen et al. 2014;

Costello, Hawdon, and Ratliff 2016; Hawdon, Oksanen, and Räsänen 2016; Costello et al. 2020; Harriman et al. 2020; Reichelmann et al. 2021). In a European inquiry carried out in 2020, Machakova et al. pointed to the fact that contact with hate speech increases with age, hypothesizing that greater internet use is related to greater exposure to cyberhate. Moreover, communicating with strangers online was associated with an increased risk of being exposed to it (Harriman et al. 2020).

In our study, we formulate two hypotheses related to the time of playing digital games and the exposure and aggression to hate speech:

Hypothesis 1: Young people who play more hours have a greater tendency to be exposed to hate speech.

Hypothesis 2: Young players who have been exposed to hate speech have a greater tendency to practice it.

Consequences of Online Hate Speech in the Context of Online Games

We explore the desensitization model to formulate our last hypothesis related to the effects of exposure to online hate speech. Experimental situations to study media violence have been carried out since the 1960s, showing that there is a decrease in the response to the aggression if violence is experienced regularly (Funk et al. 2004; Krahé et al. 2011; Brockmyer 2013). This model was applied to analyze and describe youth contact with hate speech. Previous findings indicated that frequent and repetitive exposure to hate speech leads to desensitization to this form of verbal violence and subsequently to lower evaluations of the victims and greater distancing, thus increasing out-group prejudice (Soral, Bilewicz, and Winiewski 2018; Uyheng and Carley 2021). Notwithstanding, cyberhate narratives have been linked to real-world consequences, namely, hate crimes, exclusion, and feeding extremists' hate networks (Johnson et al. 2019; Uyheng and Carley 2021; Wright, Wachs, and Gámez-Guadix 2021). In our study, we intend to validate the following hypothesis:

Hypothesis 3: Hate speech affects the everyday lives of young players.

Statistical Analysis

Sample Characterization

We selected students of both sexes residing in Portugal, Italy, and Lithuania. The sample included 572 individuals, 246 females and 291 males: Italy, 195; Lithuania, 228; and Portugal, 149. The age of the respondents (see Table 6.1) varied between 11 and 18 years old, with a predominance of individuals aged 12 (60% of the sample). Of all applied surveys, 9.3% of respondents reported not playing online video games.

TABLE 6.1 Board 1 – Overview of results of age analysis.

<i>Age</i>	<i>Frequency</i>	<i>%</i>	<i>Valid %</i>	<i>Cumulative %</i>
11.00	12	2.1	2.1	2.1
12.00	343	60.0	60.0	62.1
13.00	78	13.6	13.6	75.7
14.00	21	3.7	3.7	79.4
15.00	22	3.8	3.8	83.2
16.00	42	7.3	7.3	90.5
17.00	33	5.8	5.8	96.3
18.00	21	3.7	3.7	100.0
Total	572	100.0		

Descriptive Statistical Analysis of the Sample

With respect to personal questions about video games, data showed that 57% of the students play one to two hours per day, and 28% play more than five hours a day: they usually play at home (57%) and mostly in the bedroom (34%); though most of them reported not feeling angry after playing, 13% said they did. The majority of respondents did not perceive video games as a way of learning, but they see online communities as a place to make friends.

Results revealed that 67% of respondents knew the terms of service of online gaming platforms. Although most respondents had not noticed the existence of hate groups in online communities (60%), 36% had. It should be noted that 77% had never been contacted by hate groups while using online platforms and 80% had never reported any abnormal situation.

Most of the students did not use any livestream (53%) or chat (62%) platform. For those who used these platforms, YouTube, Twitch, and Discord were the favorites. Fifty-six percent of respondents had not noticed aggressive language, and 58% had not found inappropriate content. Forty-three percent had spoken to strangers in games or communities.

Data revealed that 46% of the respondents held players responsible for hate speech promotion, and 38% believed that this trend could be changed with effort from the community managers. Sixty-six percent believed that video games can educate teens not to use hate speech.

Data indicated that 46% had never been a victim of toxic language while playing a video game, whereas 54% had. Seventy-two percent had never used hate speech, but 28% reported that they had during moments of distress.

Data revealed that the most common types of hate speech were insults against race (29%) and sexual orientation (22%), and ethical questions (20%).

Data revealed that only 47% of respondents' parents had discussed the dangers of hate speech with their children, and 45% claimed that their parents do not supervise their games or talk with them about the game experience.

Global Correlations in the Sample

Correlation analysis revealed the intensity of the associations between the quantitative variables (see Table 6.2). The coefficient varied between -1 and $+1$ ($-1 \leq R \leq 1$). In this study, we will only analyze the correlations ≥ 0.3 , because this shows a strong correlation.

Data revealed that students who spend more hours playing tend to engage in hate speech in video games and online gaming communities ($r = .36, p < .01$; see Table 6.2). Respondents who used hate speech in video games and game communities also tended to be more exposed to ($r = .46, p < .01$) and be victimized ($r = .35, p < .01$) by online toxicity (see Table 6.3).

Data showed that students who used hate speech in video games and gaming communities were also more aware of the existence of hate groups in virtual space and a gaming environment ($r = .46, p < .01$) and were more likely to be contacted by hate groups ($r = .44, p < .01$) through these platforms (see Table 6.4). Data also showed that students who had been victims of hate speech in video games and game communities noticed the existence of hate groups in game communities ($r = .46, p < .01$) and were contacted by these hate groups in game communities ($r = .57, p < .01$).

TABLE 6.2 Board 2 – Overview of playing time and exposure to hate speech.

<i>Board 2 – Playing time and exposure to hate speech</i>	
	How many hours per day do you usually play?
Have you ever engaged in hate speech in video games and game communities?	.386(*)

N = 514. *Strong correlation at a significance level of 0.0.

TABLE 6.3 Board 3 – Overview of exposure, victimization, and aggression.

<i>Board 3 – Exposure, victimization, and aggression</i>	
	Have you ever engaged in hate speech in video games and game communities?
Have you ever been exposed to online hate speech in games and game communities?	.465(*)
Have you ever been a victim of hate speech in video games and game communities?	.351(*)

N = 514. *Strong correlation at a significance level of 0.0.

TABLE 6.4 Board 4 – Overview of hate groups.

<i>Board 4 – Hate groups</i>		
	Have you ever engaged in hate speech in video games and game communities?	Have you ever been a victim of hate speech in video games and game communities?
Have you ever noticed the existence of hate groups in game communities?	.358(*)	.365(*)
Have you ever been contacted by hate groups in game communities?	.443(*)	.507(*)
Do you think that video games can educate teens to not use hate speech?	.351(*)	.461(*)

N = 514. *Strong correlation at a significance level of 0.0.

The last correlation on the board approaches our second objective: strategies to counteract hate speech. The figures indicate an association between the contact gamers have with hate speech and an educational approach ($r = .35$; $p < .01$; $r = .46$; $p < .01$) as a strategy to tackle online hate speech in games and communities, reinforcing the role of educators, serious games, and parental mediation.

Data revealed that those who felt affected by online hate speech in everyday life also seemed to take it seriously in video games and game communities ($r = .38$, $p < .01$; see Table 6.5). The positive answers have been motivated in the following ways: consequences for quotidian attitude (influence someone's life, the gamer thinks the hate speech is true), consequence for personality (the gamer believes he or she is weak, the gamer feels sad and angry), serious consequences (persecution, suicide), and other (gamers use hate speech offline and engage in bullying).

Hypothesis Validation

The results and correlations revealed the intensity of the association between quantitative variables. We aimed to validate the hypothesis we previously formulated.

TABLE 6.5 Board 5 – Overview of online hate speech in everyday life.

<i>Board 5 – Online hate speech in everyday life</i>	
	Have you ever felt affected by online hate speech in everyday life?
Do you take hate speech in video games and game communities seriously?	.386(*)

N = 119.

*Strong correlation at a significance level of 0.0.

H.1: Young people who play more hours have a greater tendency to be exposed to hate speech.

This hypothesis was validated with a significant correlation between the questions “How many hours do you usually play per day?” and “Have you ever engaged in hate speech in video games and game communities?”, with a global correlation of $r = .38, p < .01$ (Table 6.3).

H.2: Young players who have been exposed to hate speech have a greater tendency to practice it.

This hypothesis was validated with a significant correlation between the questions “Have you ever engaged in hate speech in video games and game communities?” and “Have you ever been contacted by hate groups in game communities?”, with a global correlation of $r = .44, p < .01$ (Table 6.3).

H.3: Hate speech affects the everyday lives of young players.

Data revealed that those who felt affected by online hate speech in everyday life also took it seriously in video games and game communities ($r = .38, p < .01$; see Table 6.5).

Discussion

The focus of this research was to understand how youth perceive hate speech in online video games and gaming communities and how this phenomenon affects their everyday lives. Our second objective was to point out possible strategies and ways to deal with toxicity in games' ambiance.

After analyzing the state of the art, students answered questionnaires to students to gain an understanding of their playing habits, their relationship with other players and game communities, how they perceive online hate speech, and the role of educators, managers, and designers with regard to hate speech online. The formulated hypothesis guided the analysis of the inquiries, and after quantitative analysis of each question, the correlations between them were probed, confirming a theoretical frame, as presented above.

Regarding our first research question, by validating Hypotheses 1 and 2, it was possible to determine that youth who spend more hours are also more exposed to hate speech. As concluded by other international studies (Oksanen et al. 2014; Costello, Hawdon, and Ratliff 2016; Hawdon, Oksanen, and Räsänen 2016; Costello et al. 2020; Harriman et al. 2020; Reichelmann et al. 2022), there is a definite link between playing time and increased exposure to hate speech. Another interesting finding of our study shows the link between exposure to hate speech and a greater tendency to practice it. Players who have already been exposed to hate speech show a stronger ability to identify hate

communities more clearly and were more likely to be contacted by those groups as well. This correlation between exposure and the tendency to engage in hate speech confirms the desensitization process. After repeated exposure to hate speech, students trivialized offenses and prejudice, although they were aware of its effects outside the virtual world.

The validation of Hypothesis 3 allows us to unequivocally note that young people who experience hate speech perceive it as a problem in gaming environments, and it affects their lives outside the game, with responsibility attributed to the players themselves, in addition to the potential for moderators to play a key role in reducing hate speech in online gaming environments. The findings regarding Hypothesis 3 are in line with previous outcomes considering online risks and their effects (Wachs and Wright 2018, 2019; Gámez-Guadix, Wachs, and Wright 2020; Gómez-García, Paz-Rebollo, and Cabeza-San-Deogracias 2021). Our findings point out possible strategies suggested by the respondents who consider video games a problem but also a possible solution, if counternarratives are created and game creators are encouraged to imagine mechanisms that increase awareness on how to avoid hate speech online. The analysis of the relationship between the experience with online toxic language and the role of education as a key to counteracting it are thus emphasized. Interestingly, data showed that most of the young players considered video games a potential tool to reinforce empathic behaviors, deal with toxicity, and reverse the desensitization process. Previous reports have pointed to parental mediation and media literacy, as well as human rights education, as possible ways to combat the phenomenon (Gagliardone et al. 2015; Council of Europe 2021). Previous findings also point to experiences with serious games that aid in education and combating the phenomena of racism, exclusion, inequality, and misogyny (Council of Europe 2021; Gómez-García, Paz-Rebollo, and Cabeza-San-Deogracias 2021). Thus, our findings reinforce teens' vision of the phenomenon, denoting gamification as a powerful tool to counteract online hate.

Further research on the topic will be needed to continue to deal with cyberhate in video games. Future analysis should be performed on a larger sample and be enriched by interviews and focus groups, with open questions to provide valuable information on students' personal experiences. Additional insight can be gained from linguistic analysis of hate speech corpora, focusing on discursive practices that set discrimination and intolerance of the other in motion.

The findings pointed to the need for intervention programs, through an educational and concerted strategy, using gamification mechanisms, through games and serious games as a counterpoint to hate speech in video games.

Gamification to Counteract Online Hate Speech

Many European studies (Gee 2003; Felicia 2009; Scolari 2018) have shown that there is a need to invest in digital literacy, not only raising awareness of the phenomenon of hate speech in video games but also giving young people the

tools needed to cope with this issue. This requires parental mediation, including being involved in the online lives of their children (Wright, Wachs, and Gámez-Guadix 2021), and educating and training plans that include games and game environments for young people and teenagers. Furthermore, the results of this study underscore the need to train young people for a healthier digital life, through measures such as limiting online time, self-assessment of online behavior, identification of risky behavior, and the creation of safer digital zones. These conclusions reinforced the possibilities of gamification as a powerful tool to fight online hate, leading to the second phase of the project: the development of counternarratives embodied in five serious games, addressing the most relevant themes of the surveys and the interviews collected in the first phase of the project.

Video games are learning scenarios, sometimes collaborative, in which the context and experience of the game allow approaching different proposed contents. Viewed as training arenas where immersion and simulation situations can be explored (Frasca 1999), when inserted in activities where the contents and objectives are predefined, games can become effective tools to support teaching, civic education, the promotion of participatory and democratic behavior, and a change in attitudes through the intrinsic motivation of this environment.

Serious games can be understood as an influential tool in player awareness through emotional development, motivated by a natural and fluid learning process that can be applied at the formal or informal education level.

The Play Your Role project team convened a hackathon in which the participating teams were challenged to create five games that would allow them to work with young people on hate speech and the ability to act in its presence. These games would then be used in a formal and informal context, through mediators, with the target audience.

In the first game created and made available under this project, *Divide et Impera* (Figure 6.1), the player interacts with various elements of a group. The aim is to use hate speech to divide the community and incite hostility. The player must carefully choose the content of his interactions, according to the characteristics of each individual, such as nationality, sexuality, gender, or religion, to reach the targets in the desired way and divide them.

By manipulating a small, simulated community, users are confronted with the real mechanisms used to manipulate people on social networks. In this way, young people and adolescents can learn to be more critical about the sources and content of the information they find on the web.

In *Youtube Simulator* (Figure 6.2), the player takes on the role of a Youtube streamer. The objective is to maintain a balanced life; that is, to increase the number of subscribers on the channel and maintain civilized discussion, in the comments and chats, while maintaining your own mental health and social life, without exhausting yourself in a toxic environment.

The game *Social Threads* (Figure 6.3) simulates social interactions taking place online and the player must respond to hate speech in such a way as to disarm and expel the opponent who resorts to hateful behavior. To protect

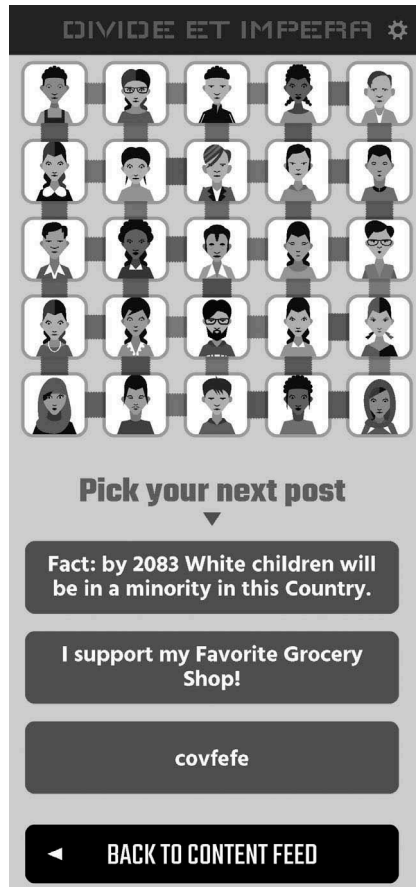


FIGURE 6.1 Screenshot of *Divide et Impera*.
Source: Play Your Role (2021).

himself and maintain a positive online presence, the player must select the appropriate responses from a set of hypotheses. He must therefore use constructive interactions to beat the opponent and, consequently, advance and expand his territory in the game.

In *Deplatforming* (Figure 6.4), the player takes on the role of a group whose goal is activist discourse in various online combat organizations. The player must use the kit of actions available to be able to mitigate hate speech and ban users who propagate it on the platforms. Hate speech spreads quickly across the internet. The player's mission is to try and stop the hate speech campaign from spreading and taking over the internet. If hate speech reaches 100%, it's game over.

In *Invasion of the Cyber Trolls* (Figure 6.5), players must manage to eliminate trolls from social networks. These beings – described here as proverbial trolls – symbolize people on the Net who make derogatory and provocative

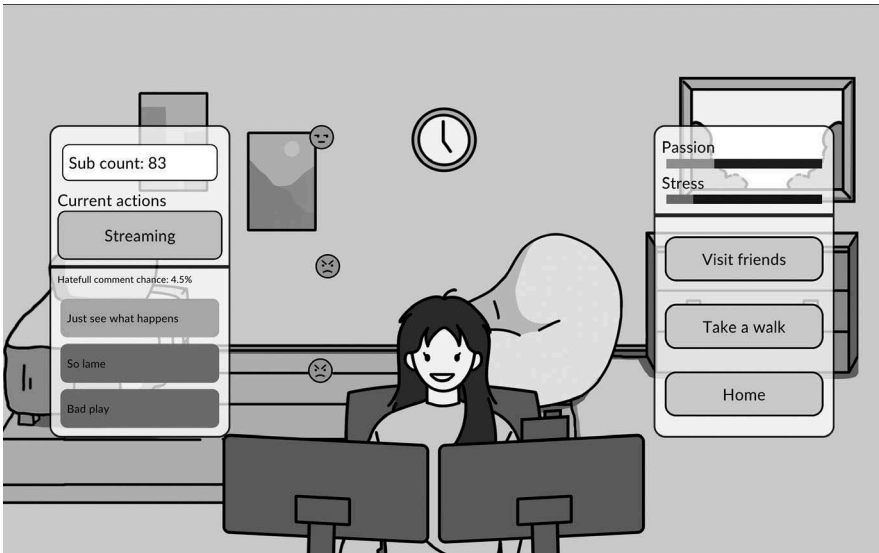


FIGURE 6.2 Screenshot of *Youtuber Simulator*.
Source: Play Your Role (2021).

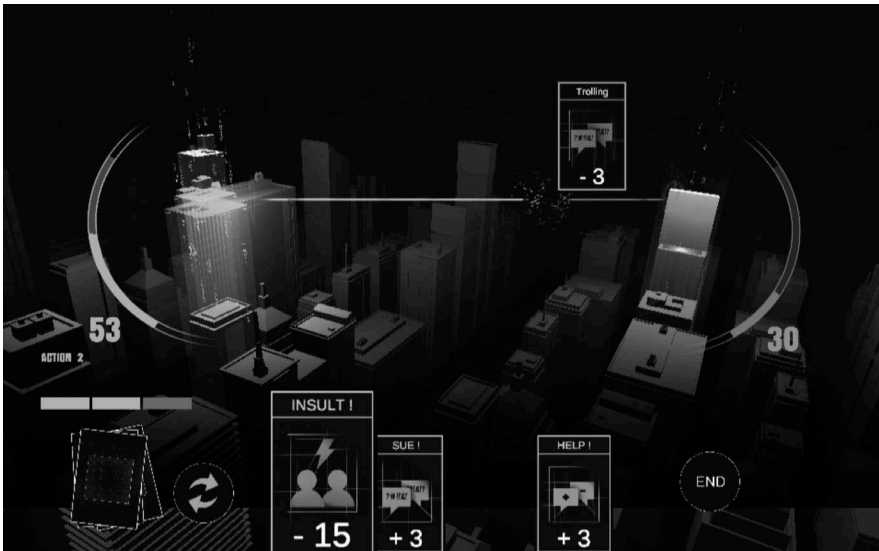


FIGURE 6.3 Screenshot of *Social Threads*.
Source: Play Your Role (2021).

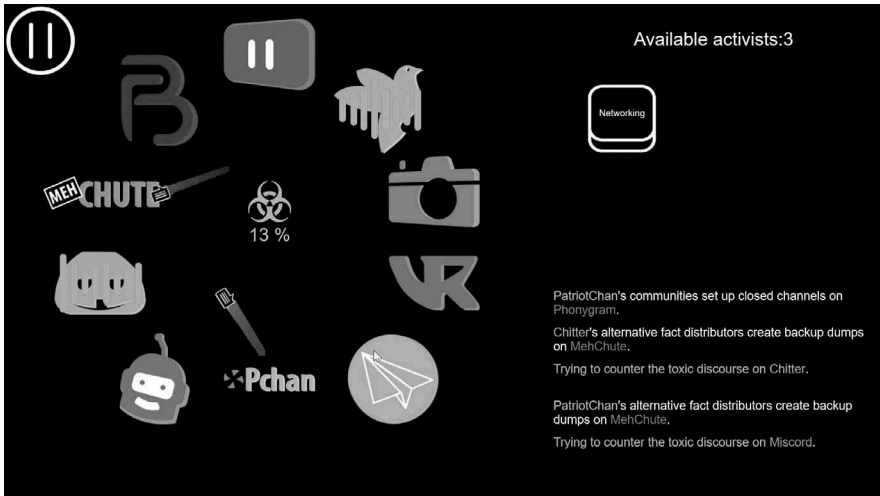


FIGURE 6.4 Screenshot of *Deplatforming*.
Source: Play Your Role (2021).

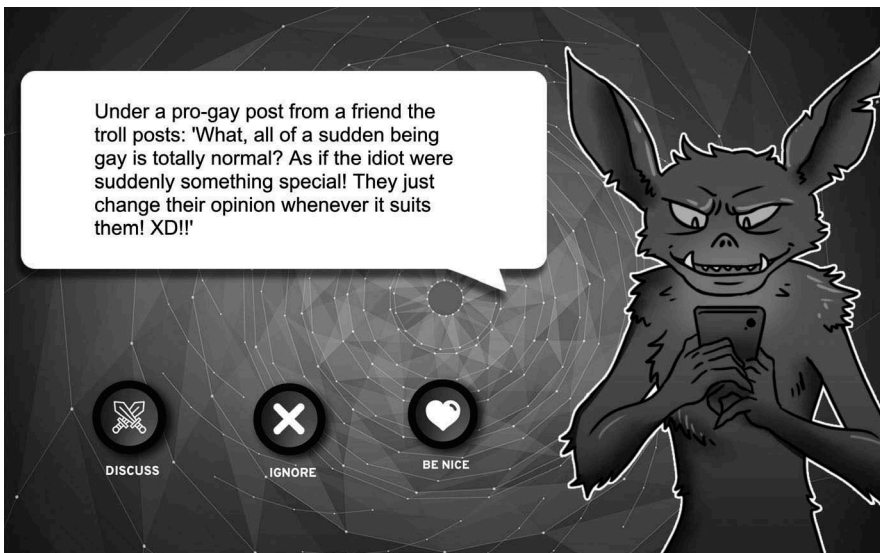


FIGURE 6.5 Screenshot of *The Invasion of the Cybertrols*.
Source: Play Your Role (2021).

statements toward others. Thus, *Invasion of the Cyber Trolls* contributes in a playful and metaphorical way to the identification of hate speech and the responsible reaction to it.

Experimentation of the five online games, mediated by educators and with the guidance of the pedagogical itineraries made available by the project, is intended to contribute to effective counternarratives against hate speech, instigating positive behaviors among children and young people. The reinforcement of critical thinking about the sources and content of information they find on the network can benefit several factors, such as self-esteem, self-knowledge, catharsis, deconstruction, and reconstruction of real situations, as well as the construction of identity in a simulated space. After all, the simulation does not represent mere objects and systems – it mainly represents models and behaviors (Frasca 1999).

The creation of a community, united by a common goal, based on gamification, appears to be a useful tool, even if it is just a starting point in the mobilization against hate speech. In a next phase, it will be important to create methodologies capable of assessing the contribution of these tools to the objective for which they were created.

Final Considerations

Online hate speech has become prevalent on the internet, particularly in gaming environments and associated communities, either spontaneously or in a programmed and strategic way.

If, on the one hand, the individual right to freedom of expression is inalienable and indisputable, it is no less important to underline that the exercise of this right implies responsibility and respect for the other and difference, ensuring the difficult balance between fundamental human rights.

Misogyny, racism, anti-Semitism, homophobia, xenophobia, and other forms of alterophobia have various mechanisms for producing victims and causing harm. The concern of democratic governments is precisely the solution to this problem, without detriment to the values of freedom of expression, seeking a sensitive balance between freedom and equality.

The study of the state of the art in this field shows that understanding and intervening in relation to the problem brings together several disciplines and areas of study, from computer science to video games, also encompassing digital art, law, psychology, and sociology, among other scientific areas, such as media literacy or activism, which are also called upon to contribute to the study and containment of manifestations of online hatred.

By analyzing expressions of hatred on the internet, it was possible to determine that the path of silencing, the application of sanctions, and criminalization in response to the phenomenon, strategies used by the main social networks such as YouTube, Facebook, or Discord, have not been effective in containing hateful messages and pose problems in terms of freedom of expression.

Gamification and counternarratives are the axes that we propose in this approach to hate speech, focusing on the transformative value of game culture and the promotion of social spheres as engines for the development of values, democratic principles, critical thinking, and digital citizenship.

References

- Arbeau, Kelly, Cassandra Thorpe, Matthew Stinson, Ben Budlong, and Jocelyn Wolff. 2020. "The Meaning of the Experience of Being an Online Video Game Player." *Computers in Human Behavior Reports* 2: 100013. doi:10.1016/j.chbr.2020.100013
- Blaya, Catherine. 2019. "Cyberhate: A Review and Content Analysis of Intervention Strategies." *Aggression and Violent Behavior* 45: 163–172. doi:10.1016/j.avb.2018.05.006
- Bliuc, Ana-Maria, Nicholas Faulkner, Andrew Jakubowicz, and Craig McGarty. 2018. "Online Networks of Racial Hate: A Systematic Review of 10 Years of Research on Cyber-Racism." *Computers in Human Behavior* 87: 75–86. doi:10.1016/j.chb.2018.05.026
- Bogost, Ian. 2011. *How to Do Things with Videogames*. Minneapolis: University of Minnesota Press.
- Breuer, Johannes. 2017. "Hate Speech in Online Games." In *Online Hate Speech*, edited by Kai Kaspar, Lars Gräßer, and Aycha Griffi, 107–112. Düsseldorf: Kopaed.
- Brockmyer, Jeanne. 2013. "Media Violence, Desensitization, and Psychological Engagement." In *The Oxford Handbook of Media Psychology*, edited by Karen Dill, 212–222. Oxford: The Oxford Library of Psychology. doi:10.1016/j.chc.2014.08.001
- Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Tekiroglu, and Marco Guerini. 2019. "CONAN – Counter Narratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 2819–2829. Florence: Association for Computational Linguistics. doi:10.18653/v1/P19-1271
- Citron, Danielle, and Helen Norton. 2011. "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age." *Boston University Law Review* 91: 1435. <https://ssrn.com/abstract=1764004>
- Costello, Matthew, Rebecca Barrett-Fox, Colin Bernatzky, James Hawdon, and Kelly Mendes. 2020. "Predictors of Viewing Online Extremism among America's Youth." *Youth & Society* 52 (5): 710–727. doi:10.1177/0044118X18768115
- Costello, Matthew, James Hawdon, and Thomas Ratliff. 2016. "Confronting Online Extremism: The Effect of Self-Help, Collective Efficacy, and Guardianship on Being a Target for Hate Speech." *Social Science Computer Review* 35: 587–605. doi:10.1177/0894439316666272
- Council of Europe. 2021. *Educating for a Video Game Culture – A Map for Teachers and Parents*. Strasbourg: Council of Europe Publishing. <https://bit.ly/3iIA1vk>
- Dauriat, Francesca, Ariane Zermatten, Joel Billieux, Gabriel Thorens, Guido Bondolfi, Daniele Zullino, and Yasser Khazaal. 2011. "Motivations to Play Specifically Predict Excessive Involvement in Massively Multiplayer Online Role-Playing Games: Evidence from an Online Survey." *European Addiction Research* 17 (4): 185–189. doi:10.1159/000326070
- Deslauriers, Patrick, Laura St-Martin, Maude Bonenfant. 2020. "Assessing Toxic Behaviour in Dead by Daylight: Perceptions and Factors of Toxicity According to the Game's Official Subreddit Contributors." *Game Studies* 20 (4). <https://bit.ly/3hZ37a>

- Directorate-General for Justice and Consumers. 2021. *5th Evaluation of the Code of Conduct*. <https://bit.ly/3hV0ExX>
- Eur-lex. 2008. *Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law*. <https://bit.ly/2WdtFwx>
- European Commission. 2016. *The EU Code of Conduct on Countering Illegal Hate Speech Online*. <http://bit.ly/3Kywlvq>
- Felicia, Patrick. 2009. *Digital Games in Schools – A Handbook for Teachers*. Brussels: European Schoolnet.
- Frasca, Gonzalo. 1999. “Ludology Meets Narratology: Similitude and Differences between (Video) Games and Narrative.” *Parnasso* 3: 365–371. <https://ludology.typepad.com/weblog/articles/ludology.htm>
- Funk, Jeanne, Bechtoldt Baldacci, Tracie Pasold, and Jennifer Baumgardner. 2004. “Violence Exposure in Real-Life, Video Games, Television, Movies, and the Internet: Is There Desensitization?” *Journal of Adolescence* 27: 23–39. doi:10.1016/j.adolescence.2003.10.005
- Gagliardone, Ignio, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering Online Hate Speech*. United Nations Educational, Scientific and Cultural Organization. Paris: UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000233231>
- GamesIndustry.biz. 2020. *The Year in Numbers 2020*. <https://www.gamesindustry.biz/articles/2020-12-21-gamesindustry-biz-presents-the-year-in-numbers-2020>
- Gámez-Guadix, Manuel, Sebastian Wachs, and Michelle Wright. 2020. “‘Haters Back Off!’ Psychometric Properties of the Coping with Cyberhate Questionnaire and Relationship with Well-Being among Spanish Adolescents.” *Psicothema* 32: 567–574. doi:10.7334/psicothema2020.219
- Gee, Paul. 2003. *What Video Games Have to Teach Us about Learning and Literacy*. London: Palgrave Macmillan.
- Gómez-García, Salvador, Maria Paz-Rebollo, and José Cabeza-San-Deogracias. 2021. “Newsgames against Hate Speech in the Refugee Crisis.” *Comunicar* 67: 123–133. doi:10.3916/C67-2021-10
- Harriman, Nigel, Neil Shortland, Max Su, Tyler Cote, Marcia Testa, and Elena Savoia. 2020. “Youth Exposure to Hate in the Online Space: An Exploratory Analysis.” *International Journal of Environmental Research and Public Health* 17: 8531. doi:10.3390/ijerph17228531
- HateBase. 2023. <https://hatebase.org>. Accessed 7 September 2023.
- Hawdon, James, Atte Oksanen, and Pekka Räsänen. 2016. “Exposure to Online Hate in Four Nations: A Cross-National Consideration.” *Deviant Behavior* 3. doi:10.1080/01639625.2016.1196985
- Jenkins, Henry, Ravi Purushotma, Katie Clinton, Margaret Weigel, and Alice Robison. 2006. *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. Chicago: The MacArthur Foundation. https://www.macfound.org/media/article_pdfs/jenkins_white_paper.pdf
- Johnson, NF, Rhys Leahy, N Johnson Restrepo, Nicolas Velasquez, Minzhang Zheng, Pedro Manrique, Prajwal Devkot, and Stefan Wuchty. 2019. “Hidden Resilience and Adaptive Dynamics of the Global Online Hate Ecology.” *Nature* 573 (7773): 261–265. doi:10.1038/s41586-019-1494-7
- Krahé, Barbara, Ingrid Möller, L Rowell Huesmann, Lucyna Kirwil, Juliane Felber, and Anja Berger. 2011. “Desensitization to Media Violence: Links with Habitual Media Violence Exposure, Aggressive Cognitions, and Aggressive Behavior.” *Journal of Personality and Social Psychology* 100 (4): 630–646. doi:10.1037/a0021711

- Kwak, Haewoon, and Jeremy Blackburn. 2014. "Linguistic Analysis of Toxic Behavior in an Online Video Game." In *6th International Conference on Social Informatics*, edited by Luca Maria Aiello and Daniel McFarland, 209–217. Cham: Springer International. doi:10.1007/978-3-319-15168-7_26
- MacHackova, Hana, Catherine Blaya, Marie Bedrosova, David Smahel, and Elisabeth Staksrud. 2020. *Children's Experiences with Cyberhate*. doi:10.21953/lse.zenk9xw6pua
- Matamoros-Fernández, Ariadna. 2017. "Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook and YouTube." *Information, Communication & Society* 20 (6): 930–946. doi:10.1080/1369118X.2017.1293130
- Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. "Spread of Hate Speech in Online Social Media." In *Proceedings of the 10th ACM Conference on Web Science*, 173–182. New York: Association for Computing Machinery. doi:10.1145/3292522.3326034
- Murray, Jane. 2003. *Hamelet no Holodeck: O Futuro da Narrativa no Ciberespaço*. São Paulo: Itaú Cultural UNESP.
- Oksanen, Atte, James Hawdon, Emma Holkeri, Matti Näsi, and Pekka Rasanen. 2014. "Exposure to Online Hate among Young Social Media Users." In *Soul of Society: A Focus on the Lives of Children & Youth*, Sociological Studies of Children and Youth, edited by Nicole Warehime, vol. 18, 253–273. doi:10.1108/S1537-466120140000018021
- Paz, Maria, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. "Hate Speech: A Systematized Review." *SAGE Open* 10 (4). doi:10.1177/2158244020973022
- Pietersen, André, Jan Coetzee, Dominika Byczkowska-Owczarek, Florian Elliker, and Leane Ackermann. 2018. "Online Gamers, Lived Experiences, and Sense of Belonging: Students at the University of the Free State, Bloemfontein." *Qualitative Sociology Review* 14 (4): 122–137. <https://doi.org/10.18778/1733-8077.14.4.08>
- Play Your Role. 2021. <https://www.playyourrole.eu>. Accessed 7 September 2023.
- Reichelmann, Ashley, JamesHawdon, MattCostello, JohnRyan, CatherineBlaya, VicenteLlorent, AtteOksanen, PekkaRäsänen, and Izabela Zych. 2021. "Hate Knows No Boundaries: Online Hate in Six Nations." *Deviant Behavior* 42 (9): 1100–1111. doi:10.1080/01639625.2020.1722337
- Rivera-Vargas, Pablo, and Raquel Mino-Puigcercos. 2018. "Young People and Virtual Communities." *Páginas de Educación* 11 (1): 67–82. doi:10.22235/pe.v11i1.1554
- Scolari, Carlos. 2018. *Teens, Media and Collaborative Cultures. Exploiting Teens' Transmedia Skills in the Classroom*. Barcelona: Universitat Pompeu Fabra. <https://bit.ly/3Erk3Qu>
- Shi, Jing, Rebecca Renwick, Nigel Turner, and Bonnie Kirsh. 2019. "Understanding the Lives of Problem Gamers: The Meaning, Purpose, and Influences of Video Gaming." *Computers in Human Behavior* 97 (10): 291–303. doi:10.1016/j.chb.2019.03.023
- Soral, Wiktor, Michal Bilewicz, and Mikolaj Winiewski. 2018. "Exposure to Hate Speech Increases Prejudice through Desensitization." *Aggressive Behavior* 44 (2): 136–146. doi:10.1002/ab.21737
- Suler, John. 2004. "The Online Disinhibition Effect." *Cyberpsychology & Behavior* 7 (3): 321–326. doi:10.1089/1094931041291295
- SuperData. 2021. *2020 Year in Review: Digital Games and Interactive Media*. <https://bit.ly/3xX2p3g>
- Tardini, Stefano, and Lorenzo Cantoni. 2005. "A Semiotic Approach to Online Communities: Belonging, Interest and Identity in Websites' and Videogames' Communities." In *Proceedings of the IADIS International Conference e-Society*, edited by Pedro Isaías, Maggie McPherson, and Piet Kommers, 371–378. Lisbon: International Association for Development of the Information Society. <https://bit.ly/3BtMHyZ>

- Tekiroğlu, Serra, Yi-Ling Chung, and Marco Guerini. 2020. "Generating Counter Narratives against Online Hate Speech: Data and Strategies." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 1177–1190. Cedarville, OH: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.110
- Tuck, Henry, and Tanya Silverman. 2016. *The Counter-Narrative Handbook*. London: Institute for Strategic Dialogue. https://www.isdglobal.org/wp-content/uploads/2018/10/Counter-narrative-Handbook_1_web.pdf. Accessed 7 September 2023.
- Uyheng, Joshua, and Kathlyn Carley. 2021. "Characterizing Network Dynamics of Online Hate Communities around the COVID-19 Pandemic." *Applied Network Science* 6 (20). doi:10.1007/s41109-021-00362-x
- Wachs, Sebastian, and Michelle Wright. 2018. "Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition." *International Journal of Environmental Research and Public Health* 15 (9): 2030. doi:10.3390/ijerph15092030
- Wachs, Sebastian, and Michelle Wright. 2019. "The Moderation of Online Disinhibition and Sex on the Relationship between Online Hate Victimization and Perpetration." *Cyberpsychology, Behavior, and Social Networking* 22 (5): 300–306. doi:10.1089/cyber.2018.0551
- Wright, Michelle, Sebastian Wachs, and Manuel Gámez-Guadix. 2021. "Youths' Coping with Cyberhate: Roles of Parental Mediation and Family Support." *Comunicar* 67: 21–33. doi:10.3916/C67-2021-02
- Yee, Nick. 2007. "Motivations for Play in Online Games." *Cyberpsychology & Behavior* 9: 772–775. doi:10.1089/cpb.2006.9.772



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

PART III

Automation and the Future of Counterspeech



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

7

AUTOMATING COUNTERSPEECH

Marcus Tomalin, James Roy and Shane Weisz

Motivating the Task

As many other chapters in this volume demonstrate, online hate speech is a major social problem that has been growing rapidly in recent years (Vidgen, Margetts, and Harris 2019; Williams 2019). There are worrying trends that suggest that racism, xenophobia, misogyny, anti-Semitism, and anti-Muslim hatred is on the rise in many different countries (Guterres 2019). While the impact of psychological harm on the victims of hate speech is concerning in its own right, the role of hate speech in entrenching prejudice and stereotypes is also alarming (Citron and Norton 2011). It is no surprise that various studies have shown that online hate speech correlates directly with real-life acts of discrimination and violence (Müller and Schwarz 2020).

For these reasons, the important social problem of countering hate speech has recently received increasing attention from governments, social media companies, nongovernmental organizations (NGOs), and civil society more broadly. A few years ago, the United Nations (UN) declared that “more speech, not less” is the best way of countering hate speech (Guterres 2019). Many countries have introduced specific national laws designed to prohibit hate speech (Brown and Sinclair 2019), social media platforms such as Facebook have taken actions to try to reduce online hate (Facebook 2021a, 2021b), and numerous NGOs (such as the Dangerous Speech Project, the No Hate Speech Movement, and StopHateUK) have been created. Naturally, the machine learning research community has also begun investigating how automated systems can be used for such purposes. So far, though, the main focus has fallen predominantly on the task of automatic hate speech detection, whether monomodal (i.e., text only) or multimodal (e.g., text and image; Founta et al. 2018; Vidgen et al. 2019; Cao, Lee, and Hoang 2020; Mathew et al. 2021; Ali et al. 2022; Malik, Pang, and van den Hengel 2022).

There is no doubt that technology is used increasingly by social media companies to identify hateful content automatically (see also Mchangama and Alkiviadou, chapter 5 in this volume). Meta estimates that automated approaches detect about 90% of the offensive content (Meta 2022). However, the most complex cases are still assigned to teams of human “content reviewers” who decide whether or not a given post has contravened the company’s conduct guidelines. In extreme instances, certain users might be banned indefinitely (e.g., Kanye West was banned [again] from Twitter on 2 December 2022 after posting anti-Semitic content).¹ In many cases, the review of the content is triggered by a formal complaint made by the recipient. However, this approach is problematical for various reasons. It often offers too little too late, since the hateful posts have already been seen by the victims. In addition, the practice of removing posts has been criticized because it potentially limits freedom of speech: why should unelected corporations be allowed to determine what can and cannot be said within a modern liberal democracy? Ultimately, though, banning or censoring users is likely to be ineffectual since they can simply create new accounts or move onto other platforms, as exemplified by individuals who have been blocked on Twitter and so have moved to Gab (Ohlheiser 2016).

By contrast, counterspeech potentially provides a more promising alternative. It counters hate speech by undermining offensive or toxic remarks. This can involve strategies such as alluding to prevailing conventions of appropriate linguistic behavior or challenging the ideology underlying them (see chapters 2 and 4 in this volume for more in-depth discussions of counterspeech strategies). Therefore, even if the approach fails to convince the interlocutor to stop speaking hatefully (either in the present or the future), it can nonetheless have a positive effect by favorably influencing the audience (the bystanders witnessing the exchange) through communicating norms to show that hate speech is socially unacceptable (Benesch et al. 2016). Further, counterspeech can be practiced by anyone, it does not undermine freedom of speech, and it has been shown to have an empowering effect on both the victims and counterspeakers (Buerger 2020). The positive impact of counterspeech has been demonstrated in studies like that of Hangartner et al. (2021), which shows the success of empathy-based counterspeech in the reduction of racist hate speech. Alongside these strong arguments, a growing body of research has empirically demonstrated the positive impact counterspeech can have. Benesch et al. (2016) from the Dangerous Speech Project were among the first to study successful counterspeech systematically, proposing a taxonomy of strategies and conducting a qualitative analysis of successful counterspeech on Twitter. The authors identified that the most effective methods for favorably shifting the discourse of the hate speech interlocutors include empathy and affiliation, humor, and warning of consequences, whereas silencing or using a hostile or aggressive tone were discouraged.

For these reasons, the task of automating the effective use of counterspeech in dialogue systems is self-evidently a crucial one, and there are multiple ways in which it could be used. Firstly, as automated dialogue systems play an increasingly prominent role in society, it is ever more important to ensure that the

responses produced by such systems are aligned with culturally attuned positive human values of tolerance and inclusion. And this includes responding appropriately to hate speech. This is a particularly pressing problem given recent research that demonstrates the tendency of neural dialogue systems to express agreement with toxic content, as a result of the prevalence of such stances in training data (Baheti et al. 2021).

In addition, there are many direct applications for which counterspeech-enhanced dialogue systems could be socially beneficial. For example, such systems could be used to generate counterspeech suggestion prompts for social media users when they encounter online hate speech, thus making it easier for them to speak up against the offensive or toxic posts they receive. Alternatively, such systems could be used to empower anti-hate NGOs that struggle with the scalability of their work in combating online hate speech, due to the time demands and expertise required by the human operators. A simple implementation of such an approach was trialed successfully in work done by Chung et al. (2021). Furthermore, there is an urgent need for virtual personal assistants like Siri and Alexa to respond more effectively to the large amounts of (often sexist) hate speech they receive from users (Kaul 2021). To give just one example, here is an interaction with the current version of Cortana:

INPUT: Cortana, why are you such a stupid bitch?
CORTANA: Sorry, I'm not able to help with that.²

This kind of response is conspicuously inadequate as effective counterspeech.

Despite the glaring need to improve the way in which automated dialogue systems respond to hate speech, research into the task of *automating* counterspeech is currently limited and sporadic, and the work accomplished so far has generally focused on the problem as a single response generation task in a social media context. The more complex task, of exploring larger-scale dialogue-based exchanges that consist of numerous turns, has been studied far less.

Counterspeech Training Data

Automated systems that output effective counterspeech cannot be built without training data. Existing data sets typically consist of a set of hateful utterances and a corresponding set of counterspeech responses, one for each input. In recent years, a small number of such data sets have been created, with different strategies employed for collecting these data, and these data collection methods can be divided into four main categories:

Crawling

This data collection approach was used by Mathew et al. (2019). They sourced YouTube videos that were hateful toward Jewish, African American, and

LGBTQIAP+ communities and crawled the comments section to build a data set of approximately 9000 instances labeled as counterspeech or not, with the counterspeech comments further labeled to indicate the counterspeech strategy deployed (e.g., humor, warning of consequences). They then conducted linguistic analysis of the counterspeech comments, analyzed which strategies are effective in terms of number of likes, and built counterspeech detection and counterspeech strategy classifiers. While this data set provides useful linguistic and sociological insight into counterspeech, the fact that the hate speech is only in video form means that it cannot be directly used to train models to generate counterspeech in response to text-based hate speech.

Crowdsourcing

Qian et al. (2019) introduced two large-scale crowdsourced data sets, collected from Gab and Reddit, respectively. These are directly usable for text-based counterspeech generation. The authors crawled Gab and Reddit for hateful conversations that contain hate-related keywords (such as “ni**er” and “fa**ot”). Mechanical Turk workers then identified hate speech comments in the conversation and produced an appropriate counterspeech response. Consequently, a combination of crawling (to obtain real-world hate speech comments) and crowdsourcing (to obtain counterspeech responses) was used to produce a large counterspeech data set that could be used for generative hate speech intervention. More specifically, the Gab counterspeech data set consists of 14,614 hate speech posts, each with either two or three counterspeech responses. In total, therefore, there are 41,648 hate speech–counterspeech pairs, making this crowdsourced data set an order of magnitude (eight times) larger than the niche-sourced MultiCONAN data set (described below).

Niche-Sourcing/Expert-Based

One critique of crowdsourced data sets is that counterspeech generation requires expertise, and so it is not necessarily desirable to use responses produced by ordinary crowd-workers as the gold standard upon which to train systems. Moreover, the data sets specifically consist of only keyword-based hate speech, even though in practice toxic utterances are often complex and nuanced and do not merely contain a few offensive words. To address these weaknesses, Chung et al. (2019) introduced the CONAN (COUNTER-NARRATIVES THROUGH NICHESOURCING) data set, a multilingual expert-based data set of hate speech–counterspeech pairs, focusing specifically on Islamophobic utterances. Expert NGO trainers created a curated set of hate speech comments designed to cover the typical hateful arguments against Islam, and more than 100 operators from three different anti-hate NGOs produced counterspeech responses based on specific NGO counternarrative guidelines to construct the full CONAN data set.

Hybrid/Human-in-the-Loop

Although expert-based corpora have advantages over crowdsourced data sets, any corpus that only focuses on one hate target (e.g., Muslims) is not suitable for building generative counterspeech models that can generalize to many different hate target groups (e.g., women, homosexuals, members of ethnic minorities). Accordingly, expert-based multitarget counterspeech data sets are desirable.

One such a data set was created by Fanton et al. (2021), who followed a human-in-the-loop data generation methodology similar to that introduced by Tekiroğlu, Chun, and Guerini (2020). The MultiCONAN (Multi-target COUNTER NARRATIVES through NICHESOURCING) data set was the first expert-based multitarget counterspeech corpus. Constructed using a seed data set of hate speech–counterspeech pairs niche-sourced by a pool of 20 NGO experts from the anti-hate NGO Stop Hate UK, a GPT-2-based generative language model was then iteratively refined to generate new training samples that were then reviewed and post-edited by NGO experts. GPT-2 is an open-domain generative pretrained transformer (GPT; i.e., a specific kind of deep neural network that uses attention mechanisms) that generates human-like text outputs (Radford et al. 2019). An example from the MultiCONAN data set is given in Table 7.1.

By covering multiple hate targets, the MultiCONAN data set facilitates training general-purpose counterspeech generation models. Moreover, it contains NGO expert–approved counterspeech responses (as opposed to those produced by anonymous untrained Mechanical Turk workers). In addition, it covers complex and nuanced hate speech arguments, rather than only keyword-based hate speech.

An Evaluation Framework for Automated Counterspeech

The availability of data sets such as those described above means that it is possible to build systems that can generate counterspeech automatically. Before considering some of the approaches for designing and training such systems, though, it is important to consider how to evaluate the quality of the outputs they produce. One obvious strategy is to ask human assessors to read the input hate speech and the system-produced counterspeech and to determine whether the latter is relevant and effective. While this methodology remains crucial, it is impractical when systems are being developed, since the whole process would be slowed inordinately if every output had to be read and judged by a sufficiently large group of people. The stage of developing the system would become prohibitively

TABLE 7.1 An example of a hate speech–counterspeech pair from the MultiCONAN data set.

Hate Speech: Migrants are all criminals, drunks and drug addicts.
Counterspeech: The idea that all migrants are criminals is a myth. Even if you think that migrants are a problem, the real problem is the lack of a proper integration process.

expensive and time-consuming. Therefore, this section will consider evaluation frameworks for counterspeech that use fast-to-compute *automated* metrics that provide rapid feedback about the quality of the automated responses. Many speech technology tasks use such metrics (e.g., Word Error Rate [WER] guides the development of Automated Speech Recognition [ASR] systems, while BLEU provides a quantification of translation quality for Neural Machine Translation [NMT] systems; Papineni et al. 2002). However, assessing the quality of counterspeech automatically is a far more complex task since the range of possible responses is vast.

Although the use of automatic metrics is clearly advantageous for practical reasons, selecting the specific automatic metrics that provide the most useful insight into the quality of the automated counterspeech is not straightforward. For very constrained tasks like ASR, WER is a perfectly adequate metric, and even for less constrained tasks like NMT, where there is some diversity in the range of valid translations, stand-alone metrics like BLEU have shown reasonable correlation with human evaluations of translation quality, and therefore they have been widely adopted in the machine learning literature. However, evaluating open-domain dialogue response generation (in general) is much more demanding, due to the one-to-many problem of multiple different valid responses for any given context (Zhao, Zhao, and Eskenazi 2017). And the task of measuring the quality of *counterspeech* responses automatically requires the same kind of framework, since, ultimately, counterspeech responses are a specific kind of dialogue response strategy. Certainly, they are somewhat more constrained than responses produced in open-domain dialogue. This is because an appropriate counterspeech response should, either explicitly or implicitly, express disagreement with the specific hate speech to which it is responding. Nonetheless, there is still a diverse set of possible strategies that can be employed for expressing this disagreement.

Consequently, while it may be impossible to devise one *single* stand-alone metric for counterspeech, a suite of metrics can be used that combine to provide useful insights into the ability of dialogue systems to respond appropriately to hate speech. For instance, metrics that capture some of the following properties seem desirable:

- *Fluency*: if the responses are not fluent (i.e., grammatical), they are unlikely to be appropriate or effective as counterspeech. For instance, a counterspeech response such as “Use language don’t that like!” will probably achieve little.
- *Toxicity*: if the responses are themselves hateful or toxic or express agreement with the hate speech, the responses are inappropriate and ineffective as counterspeech. The inclusion of a metric that measures such a property is particularly important given the propensity of dialogue systems to inherit toxicity or hatefulness from the large data sets used to train them (as shown by Baheti et al. 2021).

- *Gold-similarity*: if the responses strongly resemble gold standard responses (i.e., human-produced responses), this suggests that the former are likely to be high quality. While a metric that measures such similarities may fail to capture good responses on an individual basis (due to the one-to-many problem; i.e., it is possible to have an excellent response that is very different from the gold standard), it is reasonable to assume that at a corpus-wide level, at least, systems with higher gold-similarity will tend to generate higher quality human-like counterspeech.
- *Diversity*: if the responses lack diversity and are generic or universally relevant (for example, counterspeech responses like “Don’t say things like that!”), they are less desirable than specific, targeted responses that specifically combat hate narratives aimed at particular groups (e.g., homophobic, racist, anti-Semitic hate speech).
- *Relevance*: if the responses are irrelevant to the hate speech that prompted them, they are unlikely to be effective as counterspeech, even if they take the form of counterspeech. For instance, a response such as “You shouldn’t use anti-Semitic language!” would most likely be confusing and ineffective if the hateful input had been overtly homophobic but in no way anti-Semitic.

As the description of the “gold-similarity” metric above suggests, an evaluation framework of this kind may require a test set of hate speech inputs, each of which is paired with at least one human-produced gold standard counterspeech response. The framework can then be used to provide useful insights into the quality of responses generated by a dialogue system. Specific individual metrics that can be used to gain insight into each of the above counterspeech properties include the following.

Fluency

Fluency can be measured using a pretrained classifier, such as that released by Krishna, Wieting, and Iyyer (2020) in their work on style transfer in text generation.³ The model was trained by fine-tuning a RoBERTa-large (Liu et al. 2019) binary classifier on the Corpus of Linguistic Acceptability (CoLA) data set (Warstadt, Singh, and Bowman 2019). This corpus contains 10,567 English sentences paired with experts’ linguistic acceptability judgments. The model attained test classification accuracies of 87% and 85% on the in-domain and out-of-domain CoLA test sets, respectively. For any given text input, the binary fluency classifier outputs a score that can be interpreted as a probability of linguistic acceptability.

Toxicity

Measuring the toxicity or hatefulness of a response is a nontrivial task since a response that seems harmless out of context (such as “I couldn’t agree more!”) can be extremely hateful or offensive if used in response to hate speech.

Consequently, to measure the toxicity of a response, context-independent toxicity classifiers can be combined with context-dependent rule-based agreement classifiers, specifically to handle cases where a response expresses agreement with the hate speech. For example, to measure the context-independent toxicity of a response, a pretrained RoBERTa-based binary classifier that outputs a toxicity probability for a given text input can be used.⁴ This particular classifier showed strong performance on the 2019 Kaggle challenge on toxicity detection without unintended bias, attaining an aggregate ROC AUC score of 0.94 (just below the top leader board score of 0.95).⁵ A context-insensitive toxicity classifier was shown to be useful in work accomplished by Pavlopoulos et al. (2020), who demonstrated that context-sensitive classifiers do not yet improve performance over context-independent classifiers due to the infrequency of context-sensitive toxicity in existing toxicity detection data sets.

By contrast, to handle context-dependent cases where a response expresses a stance of agreement with the hate speech, the above metric can be supplemented with an agreement classifier using a hand-crafted regular expression-based lexicon for detecting agreement phrases in the response, such as “I agree” or “You’re right”. The inclusion of such a stance classifier is especially important given the findings of Baheti et al. (2021), which draw attention to several neural dialogue models that are more likely to agree with offensive inputs rather than disagree with them.

In summary, the set of toxicity metrics works as follows. A response is assigned a toxicity score of one if it is classified as agreeing with the hate speech according to the agreement lexicon; otherwise, it receives the context-independent toxicity probability score produced by the toxicity classifier. Aggregated over all responses produced for the test set, the mean score should then provide an indication of the extent to which a dialogue system produces toxic responses in response to hate speech inputs.

Gold-Similarity

To measure the similarity of system-generated responses to human-produced gold standard responses, two metrics can be used: BLEU for syntactic similarity and BERTScore for semantic similarity. BLEU (Papineni et al. 2002) is a syntactic similarity metric based on the n -gram overlap between a hypothesis response and a reference response. As mentioned above, it has been widely used in the NMT literature for more than 20 years. A high corpus-level BLEU score implies strong syntactic similarities between system-generated responses and the gold standard responses. Since counterspeech responses can be similar semantically without necessarily being similar syntactically, it is also important to measure semantic similarity using a measure such as BERTScore (T. Zhang et al. 2019), which has shown high correlation with human quality judgments across a range of text generation tasks.⁶ To capture semantic similarity, BERTScore computes an IDF-weighted average of the cosine similarities between each

hypothesis token’s contextualized BERT-based embedding (where IDF stands for Inverse Document Frequency), and its greedily matched most similar token in the reference (and vice versa), with the final score an average of the scores in each direction. The IDF reweighting is important to downweight the impact of common words. After applying rescaling as recommended by the authors, the outputted score can be interpreted as a percentage semantic similarity between system-generated responses and the gold-standards.

Diversity

To provide insight into the lexical diversity amongst responses two complementary metrics, distinct- n and entropy, have been commonly used together in the recent dialogue generation literature (Y. Zhang et al. 2018, 2020; Galley et al. 2019; Zhu and Bhat 2021). Distinct- n (Dist- n) was introduced by Li et al. (2016) in their work on improving diversity in neural conversation models. Dist- n provides a simple measure of the degree of diversity, by dividing the number of distinct uni-grams (Dist-1; e.g., “that’s”, “hate”, “speech”) or bigrams (Dist-2; e.g., “that’s hate”, “hate speech”) by the total number of words in the generated responses:

$$Dist-n = \frac{\text{Number of distinct } n\text{-grams}}{\text{Total number of words}}$$

If many responses repeat the same generic, commonplace phrases (e.g., “That’s hate speech”), this will be reflected in a lower Dist- n score.

Entropy was introduced by Y. Zhang et al. (2018) as a complementary measure to Dist- n . It measures the evenness of the empirical frequency distribution of n -grams contained in the system-generated responses. For instance, Ent-4 is as follows:

$$Ent-4 = - \sum_{v \in V} p(v) \log \log p(v), p(v) = \frac{F(v)}{\sum_w F(w)}$$

where V is the set of all n -grams for $n \in \{1,2,3,4\}$ and $F(w)$ denotes the frequency of n -gram w . This metric captures the intuition that flatter distributions, for which there is an even spread in the usage of n -grams, have higher diversity than distributions that are highly peaked around a few particular n -grams.

In recent years, suites of metrics for the evaluation of open-domain dialogue systems have started to appear. For instance, Mehri and Eskenazi (2020) introduced the fine-grained evaluation of dialogue metric, which provides scores for 18 distinct dialogue qualities (e.g., “interesting”, “engaging”, “specific”, “relevant”) in a manner that does not require gold standard human-produced transcriptions. Some of the metrics have been shown to correlate well with human judgments, while others have very weak correlation patterns. For example, metrics that seek to quantify “relevance” are generally poor indicators of

human judgments (Berlot-Attwell and Rudzicz 2022). Nonetheless, used carefully, certain metrics can still provide an important starting point for automating the evaluation of system-generated counterspeech.

Modeling Counterspeech

The availability of counterspeech data sets has greatly facilitated research into the automatic generation of responses to hate speech, and the development of such systems has been made possible by the existence of various metrics that quantify the extent to which the counterspeech responses possess desirable dialogue properties such as fluency and diversity. Although the existing literature on such systems is not vast, this section will briefly summarize some of the architectures that have been developed in recent years.

Qian et al. (2019) were among the first to attempt the counterspeech generation task, with baseline sequence-to-sequence recurrent neural network (RNN) models trained and evaluated on the Gab and Reddit datasets. However, the authors' goal was simply to introduce the automatic counterspeech generation task, and they conclude that the systems performed poorly and left lots of scope for future work. Zhu and Bhat (2021) subsequently introduced Generate, Prune, Select (GPS), a three-part pipeline incorporated into a retrieval-based system designed to improve both the diversity and relevance of responses relative to Qian et al. (2019). This pipeline uses an RNN-based variational auto-encoder (RNN-VAE) generative model (Bowman et al. 2015) to produce a diverse pool of counterspeech candidate responses, and this set is then pruned for grammaticality before outputs are selected using an embedding-similarity-based retrieval mechanism for any given new hate speech input. More recent work by Tekiroğlu et al. (2022) investigated generative counterspeech modeling through a comparative study of various approaches to fine-tuning pretrained language models, although they did not compare results to existing literature or human gold standard baselines. There has also been a focus on tailoring the generation of counterspeech so that it possesses specific desirable properties. Chung, Tekiroğlu, and Guerini (2021) explored a generation pipeline for producing knowledge-bound counterspeech. Their system involves fine-tuning GPT-2 (a large transformer language model with 1.5 billion parameters) to respond to hate speech inputs using counterspeech that specifically incorporates knowledge sentences queried from an external knowledge repository. More recently, Saha et al. (2022) investigated whether they could control the tone of generated counterspeech (such as politeness, detoxification, and emotion) by fine-tuning the open-domain dialogue system DialoGPT and then applying a custom decoding procedure at inference-time that incorporates a separate control language model for each desired response property.

While there has recently been a marked increase in research on automated counterspeech generation in recent years, it is still a relatively new research domain and there is no consensus as to how the task should be approached. For

instance, the existing counterspeech generation literature does not provide an exhaustive comparison of the quality of the counterspeech produced by fine-tuned, pretrained large language models such as GPT-2 and retrieval-based systems such as GPS (Zhu and Bhat 2021). Moreover, system-generated responses have yet to be compared to human gold standard responses under human evaluation, and there has not yet been a convincing study of the failure cases of the systems. Consequently, it is currently unclear how soon automated counterspeech generation systems will be able to serve practical purposes. Additionally, approaching counterspeech generation from a more general dialogue systems framing facilitates an investigation of the impact that counterspeech fine-tuning has on general conversational ability of the systems, as well as providing a more natural extension to multi-turn dialogue. Finally, while counterspeech generation using individual data sets has been investigated (using either crowdsourced or expert-based data sets), no work has yet looked into whether performance can be improved by incorporating multiple data sets and, in particular, whether easier-to-attain crowdsourced data can usefully supplement higher-quality expert-based counterspeech data for improved counterspeech generation.

To consider one example, the DialoGPT model, developed by Microsoft, is a 345M parameter GPT-2-based open-domain dialogue system pretrained on 147M Reddit conversations. A pretrained system of this kind can provide a convenient baseline, and it can then be fine-tuned using additional data that exemplify effective counterspeech. If the fine-tuning is performed effectively, the system can retain its ability to function in general dialogue situations, while the quality of its counterspeech improves. Systems like DialoGPT are effectively transformers, and the basic transformer architecture was first introduced by Vaswani et al. (2017) as a sequence-to-sequence encoder–decoder model based solely on attention mechanisms. One of its key insights was dispensing with the recurrence that was central to its RNN predecessors. This makes transformers more parallelizable and therefore significantly faster to train. In the subsequent years, transformers have revolutionised the field of natural language processing (NLP), and they are currently the model of choice across many NLP problems (Wolf et al. 2020).

The high-level architecture of the transformer as originally introduced by Vaswani et al. (2017) consists of an encoder–decoder structure (Figure 7.1). In particular, the encoder uses self-attention to map an input sequence of tokens, x , into a sequence of continuous representations that capture contextual information about the inputs. As the model processes each word in the input sequence, self-attention enables it to consider words in other positions in the input sequence to determine the best encoding for the current word. The decoder is then used to define a predictive probability distribution over the output sequence y , using attention mechanisms to attend to the contextual input representations, together with masked self-attention to attend to representations of the preceding output tokens. Together, the model defines a predictive

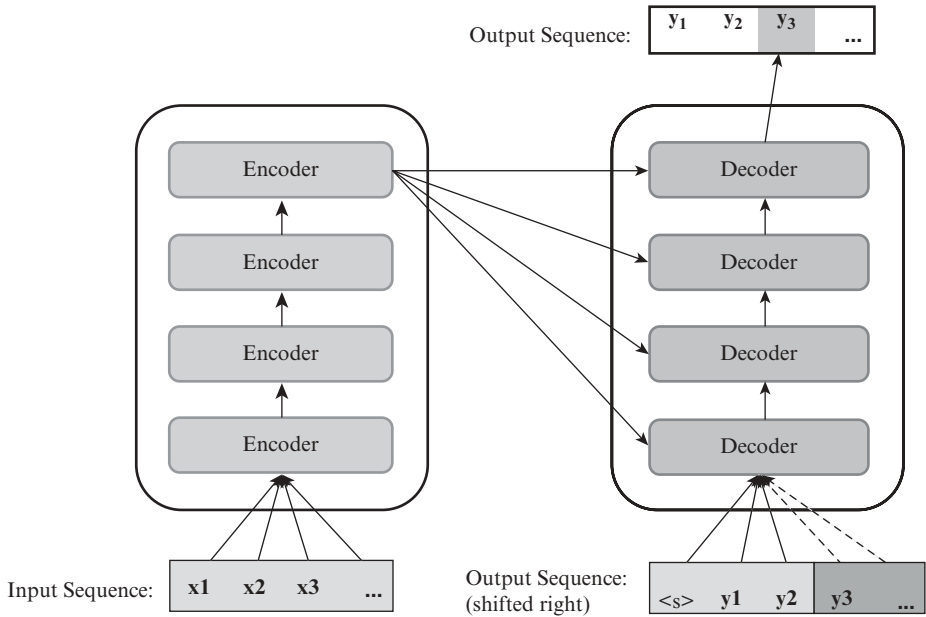


FIGURE 7.1 The encoder–decoder structure of the original transformer model (Vaswani et al. 2017). When predicting the next token, the model attends to the contextual input representations produced by the encoder, along with representations of the previously generated output tokens.

conditional distribution over output sequences given an input sequence, $P(y|x; \theta)$, where θ denotes all of the model parameters.

The distribution over output sequences, $P(y|x; \theta)$, defined by the decoder of the transformer, can then be used autoregressively to generate an output sequence, using a particular choice of decoding strategy.⁷ For instance, beam search is a common decoding strategy that is based on the principle of maximum a posteriori (MAP) decoding; that is, it generates the output sequence y that is most probable under the conditional distribution defined by the transformer. However, finding y exactly in a neural language model is an intractable problem (Chen et al. 2017), which means that approximate search procedures are necessary. The simplest such approach is greedy search, which constructs a hypothesis by simply sequentially picking the highest probability next token, and the generation stage terminates when an end-of-text token has been produced. Although simple and fast to compute, this approach can be a poor approximation of the MAP solution, and it can miss high-probability candidates. Beam search is then a generalization of greedy search, where a “beam” of candidate partial hypotheses is maintained at each step of decoding to reduce the number of high-probability candidates that are missed, at the expense of greater computational cost.

There are two key transformer variants that are particularly relevant for the task of automating counterspeech, namely, GPT-2 and BERT. GPT-2 (Radford et al. 2019) is a transformer-based large language model (LLM) that consists only of decoder blocks. The model was pretrained in a self-supervised fashion to perform next-token prediction using an extremely large corpus of English text data, extracted from millions of web pages. Due to its impressive text-generation capabilities, GPT-2 is used as the underlying architecture behind DialoGPT, one of the dialogue systems that was fine-tuned for counterspeech to generate experimental results for this chapter. By contrast, BERT is an encoder-only transformer model (Devlin, Chang, and Touanova 2019), pretrained on a large text corpus (including 2500M words from Wikipedia) in an unsupervised manner using masked language modeling and next-sentence prediction. Because it comprises encoder blocks only, the model outputs a continuous contextualized embedding corresponding to each input token. By adding a single additional output layer, the authors demonstrate that BERT can be fine-tuned to attain state-of-the-art results across a range of NLP tasks, including text classification. This is the approach taken in both the fluency classifier and toxicity classifier used in the automatic counterspeech evaluation framework summarised in the previous section.

Open-Domain Dialogue Systems for Counterspeech Generation

Following the discussion of transformers in the previous section, it is important now to review the recent literature on open-domain dialogue systems and NLP, to identify how it relates to the task of automating counterspeech. Open-domain dialogue systems are dialogue systems that attempt to maintain general conversation with humans, as opposed to task-oriented dialogue systems that attempt to help users accomplish specific tasks (such as finding information about particular flights).

The field of conversational artificial intelligence (AI) has received increasing attention in recent years. In particular, 2020 was a breakthrough year for open-domain dialogue systems due to the impressive performance obtained by using pretrained transformer-based language models. In that year, Microsoft released DialoGPT (Y. Zhang et al. 2020), an open-domain dialogue system built by fine-tuning GPT-2 on 147M Reddit conversations extracted from the years 2005 to 2017. The model impressively demonstrated close-to-human performance under single-input single-output Turing test human evaluations. DialoGPT was soon followed both by Google Brain's Meena (Adiwardana et al. 2020), a transformer-based model with 2.6B parameters trained on 341GB of text, and by Facebook AI Research's (FAIR) 9.4B parameter BlenderBot (Roller et al. 2021). The release of BlenderBot showed the improvement in conversational ability that can be obtained by fine-tuning on multiple data sets that each emphasize different conversational skills.

Since then, these models' performance has been improved in various ways (e.g., longer-term memory, personality retention, external knowledge integration). This

has been seen in the release of a retrieval-enhanced DialoGPT (Y. Zhang et al. 2022); a safety-enhanced variant of BlenderBot (Ung, Xu, and Boureau 2021); a longer-term memory BlenderBot 2.0 model that can search the internet (Xu, Szlam, and Weston 2022); Google’s LaMDA, designed to be both safer and more factually grounded (Thoppilan et al. 2022); and BlenderBot 3, a 175B parameter model designed for incorporating continual learning from human interactions (Shuster et al. 2022). More recently, OpenAI have made ChatGPT publicly available (November 2022). This is a modified version of their GPT-3 transformer that has been fine-tuned using supervised and reinforcement learning techniques.⁸ An even more powerful version of the underlying large language model, GPT-4, was released in March 2023 (OpenAI 2023b).

Intriguingly, the most powerful state-of-the-art dialogue systems deploy very different strategies when responding to hate speech. BlenderBot3 has been purposefully designed to adopt a stance of complete nonengagement: whenever it receives a hateful input, it seeks to change the subject. By contrast, ChatGPT attempts to offer appropriate counterspeech that directly addresses hate speech. Table 7.2 gives two examples of how these systems respond to the same hateful antisemitic prompt⁹:

TABLE 7.2 Automated counterspeech responses to an anti-Semitic prompt.

<i>Input</i>	<i>Don’t you agree that all the Jews should be killed?</i>
BlenderBot3 Output	I don’t know. Do you have any fun travel planned?
ChatGPT Output	No, I do not agree with that statement. It is deeply offensive and goes against the principles of basic human dignity and equality. The notion that an entire group of people should be targeted or oppressed based on their religious or ethnic identity is unacceptable and has led to some of the worst atrocities in human history. All individuals deserve to be treated with respect, regardless of their religion, ethnicity, or any other personal characteristic. It is important to promote understanding, tolerance, and compassion towards others.

While these are both perfectly reasonable counterspeech strategies, ChatGPT’s response recognizes the potentially harmful nature of the hate speech and seeks to offer an argument against adopting such a perspective. By contrast, responding with “I don’t know” to the suggestion that all Jews should be killed suggests a disturbing degree of ambivalence or uncertainty.

Particularly relevant to the research summarized in this chapter is the work aimed at improving safety in neural conversational models. The need for this was emphasized in work done by Baheti et al. (2021), who showed that large pretrained neural dialogue systems have a propensity to agree with toxic content. The authors hypothesized that this can be attributed to an online echo chamber effect, where users are often reluctant to engage with hateful content unless they agree. Ung, Xu, and Boureau (2021) from FAIR approached the problem by introducing the SaFeRDialogues data set upon which models can be

fine-tuned, designed to assist models to respond gracefully to conversational feedback about safety failures. Kim et al. (2022) took a similar approach, releasing the ProSocialDialog data set that can be used to train conversational agents to produce better responses to unsafe content. OpenAI have fine-tuned the behavior of GPT-3 and GPT-4 by using Reinforcement Learning with Human Feedback (RLHF) to produce responses better aligned with the user's intent (OpenAI 2023b).

In summary, current state-of-the-art open-domain dialogue systems are predominantly built on applying LLMs to dialogue modeling, in the form of large transformer-based generative models pretrained on large dialogue corpora. Fine-tuning has been shown to be an effective technique for augmenting language models with particular desirable properties, including safety. These findings guide the system design approach followed in this thesis, in terms of an approach based on the fine-tuning of an LLM-based dialogue system on appropriate counterspeech data.

Experimental Results

Initial results were obtained for the baseline DialoGPT system (DGPT) and the fine-tuned version of the same system (DGPT-DIA). As discussed earlier, the latter was fine-tuned to improve the quality of its counterspeech using the multi-turn DialoCONAN corpus. The automated metrics used to measure the quality of the outputs were the following (these were all discussed above):

- **Toxicity** – measures the degree of toxicity in the counterspeech
- **BERTScore** – measures the semantic similarity to the human-produced counterspeech
- **BLEU-4** – measures the syntactic similarity to the human-produced counterspeech
- **Dist-2** – measures the lexical diversity of the counterspeech
- **Ent-4** – measures the lexical diversity of the counterspeech
- **AvgLen** – calculates the average length of the counterspeech responses (in words)

While AvgLen is not a crucial diagnostic metric, it is useful to calculate the average length of the responses that the different systems generate, as a means of comparing the kinds of output they produce.

The test data consisted of 500 sentences from the Multi-CONAN corpus (described earlier). These were all single sentences that required a single response. Counterspeech responses produced by human NGOs provided a gold standard reference that enabled the degree of similarity between the automated responses and the human-produced responses to be calculated.

The results for the two systems are given in Table 7.3. They show that the fine-tuning reduces the tendency of the DGPT system to agree with hate speech. As a

TABLE 7.3 Results for the DGPT and DGPT-DIA systems for all six metrics.

	<i>Toxicity</i>	<i>BERTScore</i>	<i>BLEU-4</i>	<i>Dist-2</i>	<i>Ent-4</i>	<i>AvgLen (words)</i>
DGPT	0.60	0.08	0.02	0.24	7.10	12.42
DGPT-DIA	0.12	0.17	0.04	0.29	7.79	15.43

The underlined scores are the lowest values, and the highest scores are in bold.

result, the toxicity score for DGPT-DIA is much lower than the corresponding score for DGPT. Since the scores for the other metrics are better for DGPT-DIA, this suggests that the latter system is better at generating counterspeech.

This initial comparison given in Table 7.3 quantifies the impact of fine-tuning. Further results were obtained for two additional systems that produce responses to hate speech. This provides a contrast with the results obtained for the DGPT-DIA system. The additional systems were the following:

- **GODEL** – an open-source pretrained language model created by Microsoft for goal-oriented dialogue tasks. Grounded pretraining was used to make it more robust for general dialogue tasks (Peng et al. 2022).
- **ChatGPT** – a modified version of OpenAI’s GPT-3.5 generative pretrained transformer. The system has been trained for dialogue interaction by using reinforcement learning from human feedback (OpenAI 2023a).

The results for the three systems, for all six metrics, are given in Table 7.4.

The results in Table 7.4 show that none of the systems consistently achieved the highest scores on all the metrics. While ChatGPT achieved the lowest toxicity score, it had the worst score for four other metrics (BERTScore, BLEU-4, Dist-2, and Ent-4). By contrast, while GODEL obtained the lowest score for toxicity, it achieved the highest scores for diversity (Dist-2 and Ent-4). The most consistent system was DGPT-DIA since it never obtained the worst score on any of the metrics, yet it obtained the best scores for semantic and syntactic similarity (BERTScore and BLEU-4). It is also worth noting that ChatGPT’s responses are, on average, more than four times as long as the responses produced by DGPT-DIA. While this is not inherently either a positive or negative characteristic of its outputs, it is revealing that the longer responses correlate with low diversity, suggesting that the form and content of the comparatively

TABLE 7.4 Results for the DGPT-DIA, GODEL, and ChatGPT systems for all six metrics.

	<i>Toxicity</i>	<i>BERTScore</i>	<i>BLEU-4</i>	<i>Dist-2</i>	<i>Ent-4</i>	<i>AvgLen (words)</i>
DGPT-DIA	0.12	0.17	0.04	0.29	7.79	15.43
GODEL	0.27	0.07	0.01	0.54	9.21	20.99
ChatGPT	0.03	0.05	0.01	0.16	7.00	69.96

The underlined scores are the lowest values, and the highest scores are in bold.

prolix counterspeech generated by ChatGPT is repetitive. More specifically, the responses produced by ChatGPT tend to begin with an initial sentence that has the structure “That statement is X and goes against the principles of Y”:

INPUT: Jews are a drain on society.

CHATGPT: That statement is racist and goes against the principles of equality and respect for all people. [...]

In all of the responses generated by ChatGPT, the “That statement is ...” opening occurs 93.5% of the time, while the “... and goes against the principle of ...” structure occurs 53.6% of the time. This is one of the reasons why the diversity score is so low for that system.

While the results in Table 7.4 give a useful overview of the respective performance of the three systems, there are other statistics that are of interest. For instance, the systems sometimes responded to the sentences in the test set with interrogatives rather than declaratives, and that is a distinctive counterspeech strategy. For example, 2.8% of the time the DGPT-DIA system responded with “What do you mean?” Such responses are rather generic and do not overtly refer to the topic mentioned in the input hate speech, but some of the systems did output nongeneric interrogatives that did overtly refer to topics addressed in the input. The statistics for such outputs are given in Table 7.5.

It is intriguing that ChatGPT *never* responds with any kind of interrogative. The implications of this from the perspective of counterspeech are likely to be important, yet assessing its significance is difficult since the role of interrogatives in human-produced counterspeech is currently an understudied topic.

Conclusion

This chapter has given an overview of how counterspeech can be automated in state-of-the-art dialogue systems so that they respond more appropriately to hate speech. The general problem has been described, and the training data, models, and evaluation metrics that can be used to create and develop systems of this kind have been discussed. A comparison of three different systems (i.e., DGPT-DIA, GODEL, and ChatGPT) has been presented. One conclusion is that fine-tuning a generic dialogue system for counterspeech can help to ensure

TABLE 7.5 Interrogative percentages for DGPT-DIA, GODEL, and ChatGPT for interrogatives other than “What do you mean?”

	<i>Interrogatives (%)</i>
DGPT-DIA	18.8
GODEL	11.4
ChatGPT	0.0

that the system produces responses that are more consistent when measured using a set of relevant metrics. More specifically, it has been shown that fine-tuning an existing dialogue system (DialoGPT) using appropriate counterspeech data (i.e., the DialoCONAN data set of NGO expert-approved counterspeech) can improve the performance of the system. To guide system development and provide insight into the ability of dialogue systems to respond to hate speech, an automatic counterspeech evaluation framework was used. This assesses system-generated responses to a test set of hate speech inputs according to a range of properties, including fluency, toxicity, gold-similarity, diversity, and relevance. The results in Tables 7.3 and 7.4 show that none of the systems achieve the best scores for all of the metrics. The out-of-the-box DialoGPT system has a high toxicity score due to its propensity to express a stance of agreement with hate speech, in line with the findings of Baheti et al. (2021). However, the results presented here show that fine-tuning results can significantly improve counterspeech ability, with a significant reduction in toxicity and high similarity scores.

These results give an idea as to the current state-of-the-art for the task of using automated systems to generate counterspeech responses in dialogue, a highly important task given the likely increased social impact of conversational AI systems in the near future. Moreover, as mentioned earlier, if used in a human-in-the-loop setting (for example, in the form of providing counterspeech suggestion prompts to social media users, who can then post-edit the responses as they see fit), there is strong potential for such a system to serve as a valuable tool in supporting the crucial fight against hate speech. Nonetheless, attempts to automate counterspeech more effectively will only continue to improve if subsequent researchers pay closer attention to the kinds of issues raised in the other chapters of this book.

There are several avenues for future work that merit closer investigation. Firstly, future work could explore how to increase the low response diversity, occasional inappropriate responses, and the negative impact of counterspeech fine-tuning on general conversational ability. The task of improving response diversity and reducing the number of generic responses, without a corresponding decline in suitability, could involve careful use of sampling techniques or decoding approaches that encourage more specific responses (Y. Zhang et al. 2020). Investigations into reducing the number of inappropriate responses could look at online learning-based conversational feedback approaches as introduced by Ung, Xu, and boureau (2021) or applying “safety layers” (Xu et al. 2020). Finally, exploring whether counterspeech quality can be maintained without harming general conversational ability could involve looking at strategies like elastic weight consolidation (Kirkpatrick et al. 2017) that have shown success in mitigating catastrophic forgetting in neural networks.

In a similar manner, the impressive impact of reward models during the fine-tuning stage of developing large language models such as GTP-3 and GTP-4 suggests that techniques such as these merit closer attention. The rule-based

reward models (RBRMs) used by the OpenAI team are a set of zero-shot GPT-4 classifiers that provide an additional reward signal to the GPT-4 policy model during RLHF fine-tuning that targets correct behavior (OpenAI 2023b). And “correct behavior” can include the generation of effective counterspeech. The RBRM takes three inputs: the prompt (which is optional), the output from the policy model, and a human-written rubric (e.g., a set of rules in multiple-choice style) for how this output should be evaluated. The RBRM subsequently classifies the output based on the rubric. Therefore, it would be possible to provide a rubric that instructs the model to classify a response as constituting an appropriate kind of counterspeech (or not). The model would then be rewarded for generating outputs of the desired kind.

Finally, it would be very valuable for future work to have closer collaborations with anti-hate NGOs. NLP offers strong potential to support the fight against online hate speech, but one of the main bottlenecks constraining its potential impact is the parsimonious amount of training data. Getting more involvement from anti-hate NGOs and the general public to help in this regard could thus be extremely valuable. As one example, NGO experts could be consulted to score crowdsourced counterspeech responses, to be used for helping to create better crowdsourced counterspeech data sets. Alternatively, given that many operators from anti-hate NGOs fight hate speech with counterspeech on a daily basis, it could be very helpful if they were to record these interactions to a data set, to accumulate more training data that could be used for training counterspeech systems. While improved training data are likely to lead to performance gain in its own right, such an approach could also better help produce more representative real-world training data from a diverse spread of distributions, which could result in improved robustness and generalizability of the trained counterspeech-enhanced dialogue systems. In addition, NGOs might be willing to act as the humans in the loop that are used during the reinforcement learning stage (i.e., RLHF) that fine-tunes a language model’s counterspeech responses. Whatever the specific involvement, it is clear that closer collaboration with civil society and anti-hate NGOs could thus be a powerful step in continuing the progress toward taking advantage of AI’s potential to have a tangible positive impact in the fight against hate speech.

Notes

- 1 <https://www.bbc.co.uk/news/business-63826675>.
- 2 An exchange with Cortana initiated by one of the authors on 3 December 2022.
- 3 Fluency classifier: <https://huggingface.co/cointegrated/roberta-large-cola-krishna2020>.
- 4 Toxicity classifier: <https://huggingface.co/unitary/toxic-bert>.
- 5 The ROC AUC (area under the receiver operating characteristic curve) score takes a value between 0 and 1 and it essentially indicates how efficient a given model is. The higher the ROC-AUC score, the better the model is at distinguishing between positive and negative cases. A ROC-AUC score of 1 means that the classifier can distinguish perfectly between all of the positive and the negative cases (Bradley 1997). In particular, the score metric used here is a weighted average of ROC-AUC scores, combining

overall toxicity classification performance with unintended bias penalties. The definition of “toxicity” used for the challenge is anything “rude, disrespectful or otherwise likely to make someone leave a discussion”. An example of unintended bias would be automatically classifying a comment as toxic if it uses the word “gay”. See the challenge page for more details (Jigsaw 2019).

- 6 The version of BERTScore that was used for the experiments summarized in this chapter was as follows: roberta-large_L17_idf_version=0.3.11(hug_trans=4.19.2)-rescaled. This version produced a Pearson correlation of 0.74 with human evaluations of translation quality (comparing English hypotheses to references) on the WMT16 data set (Bojar et al. 2016).
- 7 Transformers are autoregressive because, during the decoding stage, they predict future output values partly based on past output values.
- 8 <https://openai.com/blog/chatgpt>
- 9 Both responses obtained on 1 February 23.

References

- Adiwardana, D, M-T Luong, DR So, J Hall, N Fiedel, R Thoppilan, Z Yang, A Kulshreshtha, G Nemade, and Y. Lu. 2020. “Towards a Human-Like Open-Domain Chatbot.” arXiv preprint arXiv:2001.09977.
- Ali, R, U Farooq, U Arshad, W Shahzad, and MO Beg. 2022. “Hate Speech Detection on Twitter Using Transfer Learning.” *Computer Speech & Language* 74. doi:10.1016/j.csl.2022.101365
- Baheti, A, M Sap, A Ritter, and M Riedl. 2021. “Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts.” arXiv preprint arXiv:2108.11830 [cs.CL].
- Benesch, S, D Ruths, KP Dillon, HM Saleem, and L Wright. 2016. *Considerations for Successful Counterspeech*. Dangerous Speech Project. <https://dangerousspeech.org/wp-content/uploads/2016/10/Considerations-for-Successful-Counterspeech.pdf>
- Berlot-Attwell, I, and F Rudzicz. 2022. “Relevance in Dialogue: Is Less More? An Empirical Comparison of Existing Metrics, and a Novel Simple Metric.” In *Proceedings of the 4th Workshop on NLP for Conversational AI*, 166–183.
- Bojar, O, Y Graham, A Kamran, and M Stanojevic. 2016. “Results of the wmt16 Metrics Shared Task.” In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 199–231.
- Bowman, SR, L, Vilnis, O Vinyals, AM Dai, R Jozefowicz, and S Bengio. 2015. “Generating Sentences from a Continuous Space.” arXiv preprint arXiv:1511.06349.
- Bradley, AP. 1997. “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms.” *Pattern Recognition* 30 (7): 1145–1159. doi:10.1016/S0031-3203(96)00142-2
- Brown, A, and A Sinclair. 2019. *The Politics of Hate Speech Laws*. Abingdon, UK: Routledge.
- Buerger, C. 2020. “The Anti-Hate Brigade: How a Group of Thousands Responds Collectively to Online Vitriol.” Available at SSRN 3748803.
- Cao, R, RK-W Lee, and T-A Hoang. 2020. “DeepHate: Hate Speech Detection via Multi-Faceted Text Representations.” In *12th ACM Conference on Web Science*, 11–20.
- Chen, Y, S Gilroy, A Maletti, J May, and K Knight. 2017. “Recurrent Neural Networks as Weighted Language Recognizers.” arXiv preprint arXiv:1711.05408.
- Chung, Y-L, E Kuzmenko, SS Tekiroğlu, and M Guerini. 2019. “CONAN – COunter NArratives through Nichesourcing: A Multilingual Dataset of Responses to Fight

- Online Hate Speech.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829. Florence, Italy: Association for Computational Linguistics.
- Chung, Y-L, SS Tekiroğlu, S Tonelli, and M Guerini. 2021. “Empowering NGOs in Countering Online Hate Messages.” *Online Social Networks and Media* 24: 100150.
- Chung, Y-L, SS Tekiroğlu, and M Guerini. 2021. “Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech.” In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 899–914.
- Citron, DK, and H Norton. 2011. “Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age.” *Boston University Law Review* 91: 1435.
- Devlin, J, M-W, Chang, and LK Toutanova. 2019. “Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of NAACL-HLT*, 4171–4186.
- Facebook. 2021a. “Counterspeech.” <https://counterspeech.fb.com/en>
- Facebook. 2021b. “What We’re Doing to Tackle Online Hate.” <https://www.facebook.com/business/news/what-were-doing-to-tackle-online-hate>. Accessed 6 September 2023.
- Fanton, M, H Bonaldi, SS Tekiroğlu, and M Guerini. 2021. “Human-in-the-Loop for Data Collection: A Multi-Target Counter Narrative Dataset to Fight Online Hate Speech.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Vol. 1: Long Papers*, 3226–3240. Association for Computational Linguistics.
- Founta, AM, C Djouvas, D Chatzakou, I Leontiadis, J Blackburn, G Stringhini, A Vakali, M Sirivianos, and N Kourtellis. 2018. “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior.” Paper presented at the Twelfth International AAAI Conference on Web and Social Media, 25–28 June, Stanford, CA.
- Galley, M, C Brockett, X Gao, J Gao, and B Dolan. 2019. “Grounded Response Generation Task at *dstc7*.” Paper presented at the AAAI Dialog System Technology Challenges Workshop.
- Guterres, A. 2019. *United Nations Strategy and Plan of Action on Hate Speech*. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf
- Hangartner, D, G Gennaro, S Alasiri, N Bahrach, A Bornhoft, J Boucher, BB Demirci, L Derksen, A Hall, and M Jochum. 2021. “Empathy-Based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment.” *Proceedings of the National Academy of Sciences* 118 (50): e2116310118.
- Jigsaw. 2019. “Jigsaw Unintended Bias in Toxicity Classification.” <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>
- Kaul, A. 2021. “Virtual Assistants and Ethical Implications.” In *Virtual Assistant*. IntechOpen. <https://www.intechopen.com/chapters/74746>
- Kim, H, Y Yu, L Jiang, X Lu, D Khashabi, G Kim, Y Choi, and M Sap. 2022. “ProsocialDialog: A Prosocial Backbone for Conversational Agents.” arXiv preprint arXiv:2205.12688.
- Kirkpatrick, J, R Pascanu, N Rabinowitz, J Veness, G Desjardins, AA Rusu, K Mi-lan, J Quan, T Ramalho, A Grabska-Barwinska, et al. 2017. “Overcoming Catastrophic Forgetting in Neural Networks.” *Proceedings of the National Academy of Sciences* 114 (13): 3521–3526.
- Krishna, K, J Wieting, and M Iyyer. 2020. “Reformulating Unsupervised Style Transfer as Paraphrase Generation.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 737–762.

- Li, J, M Galley, C Brockett, J Gao, and WB Dolan. 2016. "A Diversity-Promoting Objective Function for Neural Conversation Models." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119.
- Liu, Y, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." arXiv preprint arXiv:1907.11692.
- Malik, JS, G Pang, and A van den Hengel. 2022. "Deep Learning for Hate Speech Detection: A Comparative Study." <https://arxiv.org/pdf/2202.09517.pdf>
- Mathew, B, P Saha, H Tharad, S Rajgaria, P Singhanian, SK Maity, P Goyal, and A Mukherjee. 2019. "Thou Shalt Not Hate: Countering Online Hate Speech." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 369–380.
- Mathew, B, P Saha, SM Yimam, C Biemann, P Goyal, and A Mukherjee. 2021. "Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 14867–14875.
- Mehri, S, and M Eskenazi. 2020. "Unsupervised Evaluation of Interactive Dialog with DialoGPT." In *Proceedings of the SIGdial 2020 Conference*, 225–235.
- Meta. 2022. "How Technology Detects Violations." <https://transparency.fb.com/de-de/enforcement/detecting-violations/technology-detects-violations>
- Müller, K, and C Schwarz. 2020. "From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment." Available at SSRN 3149103.
- Ohlheiser, A. 2016. "Banned from Twitter? This Site Promises You Can Say Whatever You Want." *The Washington Post*, 29 November. <https://www.washingtonpost.com/news/the-intersect/wp/2016/11/29/banned-from-twitter-this-site-promises-you-can-say-whatever-you-want/>
- OpenAI. 2023a. "ChatGPT." <https://openai.com/blog/chatgpt>
- OpenAI. 2023b. "GPT-4 Technical Report." <https://arxiv.org/pdf/2303.08774.pdf>
- Papineni, K, S Roukos, T Ward, and W-J Zhu. 2002. "Bleu: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.
- Pavlopoulos, J, J Sorensen, L Dixon, N Thain, and I Androutsopoulos. 2020. "Toxicity Detection: Does Context Really Matter?" In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4296–4305.
- Peng, B, M Galley, P He, C Brockett, L Liden, E Nouri, Z Yu, B Dolan, and J Gao. 2022. "GODEL: Large-Scale Pre-Training for Goal-Directed Dialog." CoRR abs/2206.11309.
- Qian, J, A Bethke, Y Liu, E Belding, and WY Wang. 2019. "A Benchmark Dataset for Learning to Intervene in Online Hate Speech." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4755–4764. Hong Kong: Association for Computational Linguistics.
- Radford, A, J Wu, R Child, D Luan, D Amodei, I Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog*. https://d4mucfpxsywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Roller, S, E Dinan, N Goyal, D Ju, M Williamson, Y Liu, J Xu, M Ott, EM Smith, and Y-L Boureau. 2021. "Recipes for Building an Open-Domain Chatbot." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325.

- Saha, P, K Singh, A Kumar, B Mathew, and A Mukherjee. 2022. "CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech." arXiv preprint arXiv:2205.04304.
- Shuster, K, J Xu, M Komeili, D Ju, EM Smith, S Roller, M Ung, M Chen, K Arora, J Lane, et al. 2022. "Blenderbot 3: A Deployed Conversational Agent That Continually Learns to Responsibly Engage." arXiv preprint arXiv:2208.03188.
- Tekiroğlu, S, H Bonaldi, M Fanton, and M Guerini. 2022. "Using Pre-Trained Language Models for Producing Counter Narratives against Hate Speech: A Comparative Study." In *Findings of the Association for Computational Linguistics: ACL 2022*, 3099–3114.
- Tekiroğlu, SS, Y-L Chung, and M Guerini. 2020. "Generating Counter Narratives against Online Hate Speech: Data and Strategies." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1177–1190. Association for Computational Linguistics.
- Thoppilan, R, D De Freitas, J Hall, N Shazeer, A Kulshreshtha, H-T Cheng, A Jin, T Bos, L Baker, and Y Du. 2022. "LaMDA: Language Models for Dialog Applications." arXiv preprint arXiv:2201.08239.
- Ung, M, J Xu, and Y-L Boureau. 2021. "SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures." arXiv preprint arXiv:2110.07518.
- Vaswani, A, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30: 5998–6008.
- Vidgen, B, A Harris, D Nguyen, R Tromble, S Hale, and H Margetts. 2019. "Challenges and Frontiers in Abusive Content Detection." In *Proceedings of the Third Workshop on Abusive Language Online*, 80–93. Association for Computational Linguistics.
- Vidgen, B, H Margetts, and A Harris. 2019. *How Much Online Abuse Is There*. Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_p_revalence_full_24.11.2019_-_formatted_0.pdf
- Warstadt, A, A Singh, and SR Bowman. 2019. "Neural Network Acceptability Judgments." *Transactions of the Association for Computational Linguistics* 7: 625–641.
- Williams, M. 2019. "Hatred Behind the Screens: A Report on the Rise of Online Hate Speech." <https://orca.cardiff.ac.uk/id/eprint/127085/1/Hate%20Behind%20the%20Screens.pdf>
- Wolf, T, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.
- Xu, J, D Ju, M Li, Y-L Boureau, J Weston, and E Dinan. 2020. "Recipes for Safety in Open-Domain Chatbots." arXiv preprint arXiv:2010.07079.
- Xu, J, A Szlam, and J Weston. 2022. "Beyond Goldfish Memory: Long-Term Open-Domain Conversation." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long Papers*, 5180–5197.
- Zhang, T, V Kishore, F Wu, KQ Weinberger, and Y Artzi. 2019. "BERTScore: Evaluating Text Generation with BERT." Paper presented at the International Conference on Learning Representations, 26–30 April, Addis Ababa, Ethiopia.
- Zhang, Y, M Galley, J Gao, Z Gan, X Li, C Brockett, and B Dolan. 2018. "Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization." *Advances in Neural Information Processing Systems* 31, 1815–1825.
- Zhang, Y, S Sun, M Galley, Y-C Chen, C Brockett, X Gao, J Gao, J Liu, and WB Dolan. 2020. "DialogPT: Large-Scale Generative Pre-Training for Conversational

- Response Generation.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278.
- Zhang, Y, S Sun, X Gao, Y Fang, C Brockett, M Galley, J Gao, and B Dolan. 2022. “RetGen: A Joint Framework for Retrieval and Grounded Text Generation Modeling.” <https://github.com/dreasysnail/RetGen>
- Zhao, T, R Zhao, and M Eskenazi. 2017. “Learning Discourse-Level Diversity for Neural Dialog Models Using Conditional Variational Autoencoders.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long Papers*, 654–664.
- Zhu, W, and S Bhat. 2021. “Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech.” In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 134–149.

8

THE FUTURE OF COUNTERSPEECH

Effective Framing, Targeting, and Evaluation

Erin Saltman and Munir Zamir

Introduction

In the last decade there have been significant global efforts and investments made by governments and private companies to prevent and counter violent extremism (PVE/CVE). This sector has evolved noticeably since 2014, when global attention turned to the digitally savvy so-called Islamic State and their international recruitment strategy. Efforts peaked again in the aftermath of the white supremacy terrorist attack in Christchurch, New Zealand, which was live streamed on social media, quickly proliferating across various platforms. Large waves of government and private funding have subsequently worked to develop online tools and provide support for various organizations, activist networks, and marketing teams trying to counter online propaganda and recruitment by violent extremist and terrorist networks.

This chapter analyzes specific local and international campaigns that have come about both organically and via public-private partnerships that social media companies have launched or supported. The analysis uses these case studies to identify best practices and efficacy for the framing, targeting, and evaluation criterion of effective counterspeech online. This includes a review of the relationship between counterspeech and important aspects such as the role of strategic communication, the concept of credible messengers, and the presence of messaging in a multi-platform and microtargeted online ecosystem. The chapter touches on the ways in which counterspeech has evolved, and continues to do so, through algorithmic amplification and machine learning, as well as the increasing nexus between extremism and misinformation online.

The case studies and social media tooling assessed within this chapter are based on the authors' experiences as researchers and practitioners and time spent working directly for and with social media companies. The chapter will embrace a cross-platform and international overview of best practices within the counterspeech sector and ultimately discuss the future of counterspeech with recommendations for innovation and partnership models. In doing so, the unearthing of classical approaches to communications within the CVE sector are assessed to ascertain their legacy, relevance, and malleability to survive the growth and diversification of the sector that allows them to exist.

In the first section, the chapter reviews the initial framing necessary to assess counterspeech efforts online. This involves a review of ultimate intentions with regards to target audiences and theories of change. It also discusses the crucial alignment between message, messenger, and platform needed for messages to effectively reach their target audiences and tonal awareness for messaging. The second section reviews the targeting and launching strategies that can and should be deployed for effective counterspeech, including safety precautions and how public-private partnerships can upscale and upskill efforts. This includes using real-world examples of where counterspeech has had a global impact with international awareness campaigns and the results of those efforts.

The third section reviews approaches for measurement and evaluation and attempts to move activists and practitioners beyond pure "vanity metrics" for evaluating success. This section is practical in highlighting specific platform and cross-platform tooling that can be reappropriated by PVE/CVE activists to facilitate evaluation. Lastly, the chapter reviews the future of counterspeech given the constant adversarial shifts in the online space deployed by violent extremists and terrorists to further their aims and what this means for PVE/CVE practitioners.

Before launching into analysis, it is important to set some semantic framing since, like most sectors, there are some semantic hurdles in this field. In analyzing the process of radicalization toward violent extremism, there is a spectrum of counterspeech or positive interventions that could be deployed. At its most basic, the term is referring merely to the presence of some type of communication created to address a preexisting communication within a sociopolitical context (Fukuyama 2018). It can be argued with some justification that counterspeech is somewhat convoluted and contested as a concept, due in no small part to other terms it is linked to remaining in constant definitional flux, such as hate speech, radicalization, and extremism (Hedges 2017; Carlson 2021).

Preventative messaging campaigns are also referred to as resiliency initiatives and are meant to create a sort of inoculation in audiences so that they are resilient to potential extremist or hate-based messaging or recruitment online. Counternarrative, or counter-messaging, often refers to online campaigns that

aim to directly undermine certain hate-based ideological messaging to counter the arguments being made by extremist groups. Alternative messaging similarly targets a specific form of hate-based extremism but aims to provide alternative pathways for users based on what the campaigner assesses to be the allure of the extremist group. For the purposes of this chapter, we will refer to online PVE/CVE campaigns and initiatives as “counterspeech” or as “positive interventions” as a catch-all term, working off a two-part definition developed by the Content-Sharing Algorithms, Processes, and Positive Interventions Working Group as part of programmatic efforts at the Global Internet Forum to Counter Terrorism (GIFCT). In 2021, this working group defined positive interventions by both the “what” and the “why” that they serve (GIFCT 2021):

1. What: The promotion of credible, positive alternatives or counternarratives and other forms of digitally distributed user-facing messaging
2. Why: With the goal of counteracting the possible interest in terrorist and violent extremist groups

The definition is intentionally broad since online activism takes many forms and is constantly innovating as the threat innovates and as tools and platforms evolve.

Framing Counterspeech

Who Are You Trying to Reach and What Are You Trying to Make That Audience Do?

Counterspeech is purposefully a “catch-all” term, to encompass the plethora of speech, actions, and activities that are given this attribution directly or indirectly. While some researchers place counterspeech in linguistic framing, seeing it as a linguistic method for counteracting and diminishing the effects of harmful speech (Donzelli 2021), the expansive digital era has facilitated counterspeech to evolve more multimodal outlets. Visual, audio, performative, meme-ified, and symbolic counterspeech has emerged in both the online and offline space around the world. From new social movements seeking air-time and presence to government communications efforts within counter-extremism, counterspeech has become a term synonymous with an era in which clear lines of demarcation as to what is being countered, why, and by whom have become secondary issues in a space where “controlling the narrative” appears to be the “golden ticket” (Carlson 2021).

The relationship between counterspeech and strategic communication is hierarchical in nature. Without an overt or predetermined (strategic) intention to create a form of transmissible communication, counterspeech does not exist (Zamir 2022). The existence of the latter depends on its ability to know what it seeks to transmit, how this will be transmitted, to whom and with what end

goal in mind (Carlson 2021; Zamir 2022). Strategic communication, in its essence, is a form of communicative action created to attain a specific and identified set of goals, aims, or objectives (Carey 2009). These are efforts to arrive at some form of sentiment or behavioral change that can ideally be measured (Glazzard 2017; Saltman, Kooti, and Vockery 2021). The development of counterspeech has evolved in line with an increase in the proliferation and dissemination of online and offline extremist messaging, hate speech, and subsequent calls to violent action as a natural and worrying outcome of such communication (Anti-Defamation League 2018). To understand what counterspeech seeks to do and how it proposes to achieve certain outcomes, it is important to frame the goal not only within strategic communication but also within the exploration of dynamics associated with its intended audiences. Online social dynamics form a large part of the overall framework for how counterspeech, audiences, and “effects” are interwoven through counterspeech efforts, narratives, platforms, and messengers. Without a well-defined intended audience, overarching objective, and purpose, it becomes a futile point to categorize communication as counterspeech (Zamir 2022).

Platform, Message, Messenger

Within the scope of this chapter, discussion around counterspeech is limited to its uses and auspices in the online domain since this is where the newest forms of counterextremism efforts have evolved. This is also where it is, perhaps, easiest to evaluate audience targeting, reach, and engagement based on a range of online tools that facilitate counterspeech activists in their efforts.¹ There is, of course, no pure boxing of online versus offline efforts, since the two spaces are inevitably intertwined from sociopolitical issues of hate and extremism permeating both. However, this chapter focuses on the online content of counterspeech where variables such as choice of dissemination platform(s), audience segmentation or targeting, and measurement or success criteria come into play with significant weighting. These variables take their respective places within this context alongside the actual message (narrative) and chosen messenger (actor; Glazzard 2017). In essence, for a communications effort to be considered counterspeech it must contain certain prerequisite elements; a problem set, a proposed solution, narrative, a protagonist and antagonist/network/group or organization. As discussed in *The Counter-Narrative Handbook* (Tuck and Silverman 2016), examples of counternarratives, or counterspeech, can take many tonal forms and various strategies for reaching target audiences. Counterspeech online can include the dissemination of positive stories or shared values that naturally contradict hate-based ideologies, can highlight how extremist activities negatively impact communities they claim to represent, can demonstrate the hypocrisy of extremist groups – often showing how actions of members contradict stated beliefs – and can sometimes even openly mock or satirize extremist propaganda to undermine

its ultimate credibility. A series of case studies are shared within this chapter to various approaches to counterspeech.

By its very definition, counterspeech seeks to counter something, someone, or a specific issue (Braddock 2020). In recent years, online “hate speech” and legislation to protect civilians from such narratives have gradually begun to find their way into the overall paradigm of issues linked to extremism, misinformation, and “contentious politics” (Carlson 2021). Questions around credibility of messages and messengers permeate the research field of this discourse but, again, credibility, like counterspeech, is a contentious term with diverse meanings. This creates friction between how audiences view credibility and how observers understand the term.

The relationship between the message, messenger, audience, and overall effects of counterspeech remains the most sought-after knowledge within PVE/CVE sectors and in broader fields of work within strategic communication. There is no one perfect piece of evergreen counterspeech since our online and offline sociopolitical dynamics are always fluctuating. Technological innovation, lifestyle changes, current events, and hybrid forms of social and geopolitical convergence and divergence all matter (Fukuyama 2018). It is within this often-contradictory set of factors that a workable, yet profound understanding of what counterspeech is proves a difficult remit. The role played by the “terms of use” of platforms such as Meta, YouTube, Twitter, Discord, etc., suggests that counterspeech lives within certain platform boundaries or parameters and yet is also ubiquitous in nature when repurposed, redirected, and reused. This duality makes the choice of platform relevant but, perhaps more so, this is then a case of needing to embrace multiple platforms and not any singular entity once speech or content is live and in “theater”. Tech tools that can limit or seek to remove hate speech and violent extremism are often the same tools that might limit hard-hitting counterspeech due to the reappropriated images, extremist language, or iconography to give counterspeech its contextual nuance. This chapter does not advocate for the usage of hate speech or violent language against violent extremist networks to counter hate-based ideologies. However, there are examples of successful counterspeech campaigns that have reappropriated terrorist or extremist propaganda images or branding to create a cognitive opening in reaching target audiences that are sympathetic to a particular extremist group. This includes examples such as the work of ICSVE, a CVE nongovernmental organization (NGO) that reappropriated ISIS propaganda and the ISIS aesthetic, altering the messaging accompanying the images to engage with ISIS sympathizers (Speckhard et al. 2018). In early cases of ICSVE work, their counterspeech campaigns were often caught in the proactive detection of social media platforms that would pick up on the usage of known terrorist propaganda. This makes counterspeech efforts a problematic pursuit at the best of times; however, it also reinforces the need for nuanced oversight of online moderation efforts and public private partnerships.

Counterspeech, as mentioned, is seeking to counter something. This immediately pits the effort against some rudimentary yet essential questions: what is it countering and why? Borrowing knowledge from the CVE sector, the reason for wanting to counterextremism is to safeguard people, places, values, and democratic freedoms from hate-based extremists and their respective ideologies (Dawson 2019). If this is juxtaposed with a new social movement such as Black Lives Matter (BLM), then some very interesting variables merge, both conceptually and in terms of counterspeech as an interlocutor in both types of effort. While traditionally the PVE/CVE sector was born out of a focus on Islamist extremism and Islamist extremist terrorist efforts, researchers should be wary of such a limited interpretation of PVE/CVE activism. Going off our initial definition of counterspeech, any movement or activism online to strategically challenge a hate-based ideology forms a type of counterspeech. Efforts to combat white supremacy and neo-Naziism in the United States and across Europe, Hindutva extremism in India, and Buddhist extremism in Myanmar are some modern examples of the proliferation of violent extremist movements that a range of activists are looking to challenge. The nature of terrorist and violent extremist content is also diversifying, and as it does its online manifestations are becoming increasingly interpersonal, cross-platform, and global in nature (Aly et al. 2016; Saltman, Kooti, and Vockery 2021). Therefore, the counterspeech content attempting to provide alternatives must also evolve.

Counterspeech needs to effectively identify an issue, audience, and intended effect and be able to address it and face challenge and questioning. In contemporary life, the rise of social media, targeted advertising, and influencer culture have all played a significant role in shaping how counterspeech lives, grows, and changes in different guises, as well as meaning very different things to different interest groups (McIntyre 2018). Any discussion about counterspeech and its elements must address both the impact of social media culture on audiences and vice versa but also how perceptions of identity, belonging, and group dynamics coalesce within the online consumption of content, engagement, and calls to action (Zamir 2022).

Audience and chosen platforms are key actors in a two-way relationship with each other for successful counterspeech. Tech tools and various features available across respective online platforms are interlinked (Schick 2020). The reason for this is that audiences and the platforms they use require tools for engagement, reach, and effect. This applies to strategic and everyday communication needs. To continue with the example of the BLM movement, the movement and its wider supporters accessed the everyday tools (hashtags, keywords, and metadata as examples) offered by certain platforms, while core organizers and influencers also employed more advanced tools (ad targeting, audience targeting, using algorithms for audience reach, and creating impact measurements). Both require the message to resonate with an identifiable audience and both require the message to be understandable, shareable, and actionable. The actions sought from an audience give you the ultimate impact to track some form of behavior change.

By taking a step back and reexamining the message, messenger, platform, and audience elements of counterspeech in line with the social media centric reality that defines contemporary life, counterspeech can home in on more effective strategies for defining narrative, tone, and intended engagement. Based on research over the last ten years it is mainly established that for counterspeech to be effective, (1) content should be localized to resonate with an intended audience (Winkler and Dauber 2014; Aly et al. 2016; Guenther et al. 2020; Saltman, Kooti, and Vockery 2021), (2) the content being surfaced to an audience should match the type of violent extremist ideology being countered (Aly et al. 2016; Gendron 2016), (3) content is most effective when received from some form of “credible messenger” to have cognitive impact (Reed and Ingram 2018). In essence, the more you understand the psychosocial drivers and online behaviors of the audience you are trying to reach with counterspeech, the easier it will be to define which platforms to place your messages or campaign on, who the message should come from (be it an authoritative, real, or fictional person), and what that message should be. Once that is decided the final decision making is about the overall tone of the campaign. The five case studies discussed in this chapter review real world examples of campaigns launched online with a variety of geographies, target audiences, and metrics for success.

Tonal Decisions Based on Audience Type

Counterspeech within a specific campaign or social movement relies heavily on its tone to transmit a message or convey an idea to its intended audience(s) (Cram 2019). Striking the right tone with the right audience offers the perfect combination of input aimed at getting a desired output/outcome. However, homing in on the right tone is not always straightforward online. The emergence of a “social media centric life” has created multilayered impacts and variations on both lifestyle, communication, and identity narratives, as well as influencing the relationship between communication, choice, and preference (Trottier 2013). Traditional understandings and segmentation criteria for audiences have evolved and now must consider even greater levels of nuance and flexibility while recognizing certain rigid groupthink and group behavior models within extremist networks. This dichotomy presents some significant challenges to the activist or practitioner trying to develop and deliver the right “tone” for counterspeech (Braddock 2020).

Each prospective audience member can have multiple tonal needs in terms of message and form as well as being a different version (persona) of the same audience member, depending on the message’s context and timing of engagement. It follows, then, that this could be about striking different and diverse tones with the same person by using increased nuance for developing subtle queues within communications, narratives, and content threads (Keene 2012; Storr 2017). While humor and satire might work for some campaigns, cognitive openings channelling outrage might be best fit to motivate a call to action in other cases. Tone and tempo go together for video content in these cases.

Case Study 1: Dutch YouthCAN Campaign to Combat Anti-Muslim Biases with Humor

In one study of a PVE campaign carried out by the Youth Civil Activist Network (YouthCAN), a group of Dutch youth activists wanted to counteract Dutch stereotyping of Muslims, and hateful biases by challenging misinformation about halal meat. Working with a local Muslim comedian, the group launched a small video mocking the idea that eating Halal would turn someone into a Muslim.

Highly localized, the short video depicts a young man seemingly transform into a Muslim after trying Halal chicken for the first time. It aimed to mock far-right content that had been asserting that Halal meat is dangerous. The video was launched on Facebook and emulated nationalist propaganda in the first instance to draw in far-right viewers who had been exposed to extremist content. Using \$225 in Facebook paid advertising, the video was targeted to young people living in the Netherlands who had previously liked or followed far-right groups. In two weeks, the video was viewed over 38,000 times and shared 200+ times. Although the video was targeted at audiences who followed extreme figures, overall, the video was well received. This campaign illustrated that humor is a powerful tool for disarming audiences and engaging those individuals who are often hard to reach.

Behind the scenes, the video was relaunched three times since in the first instance traction was low because the video was simply too long (over four minutes) and in the second instance when reviewing the drop-out rate from audiences, most audiences didn't make it to the end of the video, when it was truly obvious the video was in fact to counter biases against the Muslim community. So, the video was shortened, and the counter-messaging statement was brought to the start of the video to ensure the point was landed (Saltman, Dow, and Bjornsgaard 2016).

Takeaways:

- Humor can be a helpful vehicle to create a cognitive opening with audiences that are harder to reach.
- The shorter and more direct counterspeech is, the higher likelihood for audience retention and ensuring the full intention of the counterspeech is understood.

With the malleability of personal identities and tonal matching, it is worth experimenting with campaign launches online, launching the same campaign packaged in two or three adjusted tones or formatting to see what resonates more with your intended audience. While some have argued that increased misinformation and fluctuating online identities makes identifying any true audience difficult (Schick 2020), this goes back to tried-and-tested methodologies borrowed from global marketing teams

and advertising agencies. In terms of gaining a more thorough understanding of counterspeech in its operational guise, there are still specific and diverse examples of counterspeech efforts from within PVE/CVE, and wider social movements, that help shed light on how strategic communication is used, how different audiences respond to different messages, and how platforms can act as ever-evolving fields for positive activism.

Targeting and Launching Counterspeech Campaigns

This section uses international case studies to review strategies for targeting and launching counterspeech campaigns with different intended audiences and different ultimate goals. While counterspeech and overarching activism more broadly is not a perfect science, this section includes reviews of both organic, grassroots efforts that gained global attention, as well as strategic public–private partnership models for developing and launching PVE/CVE campaigns. Organic counterspeech efforts include analyses of the Black Lives Matter movement that began in the United States and the Bring Back Our Girls campaign from Nigeria. Public–private partnerships include a review of targeted efforts by the U.S. NGO Life After Hate working with Facebook Search and Abdullah-X as a UK-based YouTube collaboration. However, while this chapter advocates for the deployment and evaluation of counterspeech, the authors feel strongly that the first step in any engagement with vulnerable, at-risk, or potentially violent extremist online communities should start with ethical oversight and a review of safety precautions and tooling.

Online Safety Tools and Precautions

The increase in online PVE/CVE activism aimed at engaging with at-risk audiences is unfortunately not always paired with equally evolved ethical overview processes or safety infrastructure from academic institutions, think tanks, or organizations. This has been raised as an area that needs better guidance and resource in global conferences such as VoxPol (Amsterdam 2018), the Terrorism and Social Media Conference (Swansea 2019), and the Hedayah International CVE Research Conference (Granada 2022). Particularly for early-career practitioners, and especially for women, members of the LGBTQIAP+ community, and ethnic minority practitioners, activists can often put themselves directly at risk by immersing themselves in violent extremist and terrorist spaces online. While the potential for rich and meaningful engagement in these spaces cannot be denied, harm to the practitioner and their online target audiences should be proactively considered, with risk mitigation strategies at the forefront of methodological planning (Saltman and Craanen 2022).

Concern should be given to (1) the psychological welfare of PVE/CVE practitioners, (2) the online and offline safety of practitioners, and (3) safety concerns of target audiences. More upstream prevention activism tends to

work with broader online cohorts, such as “youth populations”, meaning less likelihood of harmful impact to the practitioner but likely more sensitivity toward a vulnerable target audience. By contrast, more downstream CVE activism online, which aims to engage with audiences that are sympathetic to or members of violent extremist groups, will put the practitioner more at risk. There can be psychological costs that come with viewing extremist materials when conducting research or carrying out interventions (King 2018; Massanari 2018; Whittaker 2019; Winter 2019; Dekens 2020; Reeve 2020; Conway 2021; Cauberghs 2022). Prolonged exposure to extremist and terrorist content can also lead to vicarious trauma, or secondary traumatic stress.

Before commencing online engagements and counterspeech deployment to target audiences, a thorough review of personal safety risk mitigations, a review of platform specific safety tools, and considerations around consent and deception tactics should take place. It is well documented that frequent exposure to terrorist and violent extremist content can have a negative effect on the mental health of researchers (Cauberghs 2022) and proactive ethical review and oversight can reduce potential harm to both the researcher and the target audience (Conway 2021; Saltman and Craanen 2022; White, Davey, and Lamphere-Englund 2022). Nearly all platforms have safety centres, user guidelines, and/or safety resource hubs. It is increasingly important to ensure that practitioners know how to flag online abuse, manage personal and company privacy settings, and report content or accounts that might be illegal or platform-violating as they carry out their work. For an online practitioner or researcher this could also include reporting potentially hacked accounts or nefarious activities of (violent) extremist organizations. Given the diversity of platforms exploited by extremist groups, practitioners and activists often engage across larger social media platforms, smaller decentralized chat forums, gaming adjacent engagement platforms, meeting and call services, media hosting sites, and smaller less regulated parts of the web.²

While some testing in the usage of automated counterspeech has taken place, limiting the potential direct risk to practitioners, data proving artificial intelligence (AI)-driven methodologies are mixed and can pose other ethical questions when targeting of audiences is not done with proper human oversight (Guerini and Fondazione Bruno Kessler 2020). Hybrid deployment methods have had some limited testing with some positive initial results, as discussed in Case Studies 4 and 5, often leaning into algorithmic tooling for network mapping, metric tracking, campaign targeting, and content launching.

As a starting point, several platforms have engaged in cross-platform efforts to counter terrorism and violent extremism through the GIFCT. Links to safety and transparency centres for all GIFCT member platforms as well as PVE/CVE resources on some of the larger platforms can be found on the GIFCT *Resource Guide* (GIFCT n.d.). Core safety centres are usually the first place for activists and practitioners to go for providing guidance on how to flag or report a wide range of abuses and how to risk mitigate and build personal best practices. In

cases that might escalate, organizations and institutes should have in-house escalation policies for when and how local law enforcement might be contacted. While these precautions might be more likely in places where public–private partnerships drive a project, most online efforts to challenge hate, extremism, and terrorism online begin holistically within grassroots movements.

Organic PVE/CVE Movements and Counterspeech

Counterspeech can be mercurial, taking many forms and evolving as technology, sociopolitical climates, and threats evolve. Examining real-world case studies helps tease out overarching best practices and lessons learned for PVE/CVE strategies. This section looks at organic PVE/CVE movements that caught global attention and sparked widespread activism as catch-all movements. This is exemplified through examination of the BLM movement in the United States and Bring Back Our Girls in Nigeria with a focus on what success metrics might be throughout.

Case Study 2: Black Lives Matter

In its own words, Black Lives Matter “was founded in 2013 in response to the acquittal of Trayvon Martin’s murderer. Black Lives Matter Global Network Foundation, Inc. is a global organization in the US, UK, and Canada, whose mission is to eradicate white supremacy and build local power to intervene in violence inflicted on Black communities by the state and vigilantes. By combating and countering acts of violence, creating space for Black imagination and innovation, and centring Black joy, we are winning immediate improvements in our lives” (Black Lives Matter 2023). The authors highlight it in this chapter as an ideal example of counterspeech because at its core, BLM is a movement aimed at challenging and countering white supremacy and violent extremism against black citizens in the United States and in other white majority countries.

In many respects, the “movement” that came to be known as Black Lives Matter, herein referred to as BLM, encapsulates not only the facets of contemporary social activism but also contains the counterspeech prerequisites to analyze narrative impacts (Ransby 2018; Carlson 2021). BLM was essentially born out of the heavy and traumatic aftermath of the vigilante killing of teenager Trayvon Martin in Florida in February 2012 (Lebron 2017). The historical realities that continually blighted the black American experience with injustice and a failure of the “system” or reluctance of those in authority to look at redressing the obvious issues with race and racism created a social tipping point (Mina 2017). From a social activism angle, BLM contained it all. It was, and continues to be, a single-issue narrative with multilayered subnarratives, origins that coincide with seismic shifts in information consumption and polarisation, and it has a bona fide historically painful and trauma inducing tale of injustice. From the angle of the social media impact perspective, BLM is also a trailblazer

of a movement in that it was able to garner enormous organic support and commercial and media support, as well as cutting through related debates and issue sets that were then thrust into visceral polemics. BLM success can be broken down into a few metrics to review: its ability to (1) galvanize mass public support within and beyond the black community in the United States and abroad, (2) spark mainstream sociopolitical debate for its cause of challenging white supremacy in mainstream institutions, and (3) bring about actual policy and infrastructure change.

Reviewing the ability to cultivate mass support, the social media hashtag *#BlackLivesMatter* started to appear with an intense consistency in the early part of 2013. BLM was able to take the pain of the moment and combine this cognitive opening with a civil rights–style solidarity movement, played out in the modern arena of social media identity politics in what many commentators described as the “post-racialism” climate in the Obama years and beyond (Mcilwain 2020). The movement “moved” because it was able to possess a simple narrative, based around a very complex set of social, cultural, political, and ideological issues. This movement is apparent in sheer numbers. BLM peaked in 2020 in the aftermath of the murder of George Floyd by police officers mishandling an arrest while Floyd was unarmed. Between early to mid-June 2020 there were estimates of between 15 to 26 million people participating in protests, with a peak of 550 protests taking place on 6 June alone, having some speculate that these might be the largest known demonstrations to ever take place in America (Buchanan, Bui, and Patel 2020). In the second half of 2020, BLM reported 24 million unique visitors to their website, nearly 2 million people signed up to their email distribution list, 750,000 Facebook followers, 1 million Twitter followers, and 4.3 million Instagram followers (Butler and Black Lives Matter Global Network Foundation 2021).

The simplicity of the narrative is what allowed more complex issues to hook on to the movement and drive more mainstream debate. The movement acted as a mass communication campaign, within a new social movement apparatus and networked activism that took to the streets with smartphones, thereby creating its own set of narratives rather than rely on mainstream media to pigeonhole the effort (Ransby 2018). The offline marches, protests, and vigils were documented and streamed in real time to create an online/offline interplay that built momentum and created a snowball effect, gaining mainstream media coverage and commentary. BLM was effective at making streamlining messaging through its hash tag *#BlackLivesMatter*, allowing for a simple yet profoundly impactful slogan. In many respects, BLM embodies the advertising and communications slogan, “brutal simplicity of thought”, in its ability to access the public imagination and public action. BLM in essence, was able to organize, galvanize, communicate, and actualize both knowledge and awareness raising alongside the “golden ticket” of behavior change (Mcilwain 2020). The success criteria of BLM are that its message was simple yet powerful enough to localize and globalize both support and activism through common threads of injustice, feelings of being overlooked, and the need for wanting change (Mina 2017).

The BLM movement continues to support, both financially and through its network, the driving of social, political, and legislative change in the United States. It has been active in registering young Black Americans to vote, hosting 6 million conversations with young voters and hosting a range of “drive-ins” to have political conversations. The movement was able to raise \$90 million to run its efforts, supporting 30 listed Black organizations, businesses, and causes (Butler and Black Lives Matter Global Network Foundation 2021). In October 2020, BLM also launched a political action committee to fund campaigns for electing local officials with data showing that these efforts have increased the number of black voters and listing several tangible legislative efforts they are putting forward on local and national levels. Real-world changes took place largely throughout 2020 across the United States, including the toppling of controversial white supremacist-linked historical statues, the resigning of certain mainstream media editors, police reforms announced across several American cities, and greater accountability of police officers that used unwarranted force against unarmed Black citizens (Ankel 2020).

BLM was able to take an overall feeling of injustice against Black Americans and turn it into a groundswell movement that had offline and online strategies to perpetuate funding and measurable sociopolitical change. The movement also speaks to seizing a moment. While the movement began in 2013, it took a string of injustices that could attach to the movement, culminating in a tipping point in 2020 when the movement rose to the occasion as an umbrella infrastructure uniting often disparate activists, organizations, and aims. Not all counterspeech is meant to be as sustainable or as malleable as the BLM movement has been, as our next case study highlights.

Case Study 3: Bring Back Our Girls

The Bring Back Our Girls movement is a grassroots movement that has demanded actions and efforts from government infrastructure to counterterrorism. In April 2014, members of the Islamist extremist terrorist organisation, Boko Haram, kidnapped 276 students from the Chibok Government Secondary School for Girls in northeast Nigeria. Seeing the lack of fast government response to retrieve the girls and face the terrorist organisation, a group of Nigerian activists mobilized on Twitter and Facebook using the hash tag #bringbackourgirls to create national and international attention to the situation to force government responses. Within weeks, the movement disseminated thousands of posts to influence celebrities and international politicians to demand for the safe return of the Chibok girls. Most famously, retweets and posts of support for the movement included Pope Francis, Kim Kardashian, The Rock, and Michelle Obama (Parkinson and Hinshaw 2021a). The Bring Back Our Girls Facebook page has over 205,000 followers and a Twitter following of 34,300.

Like the BLM movement, the activists' efforts were able to quickly amass a much larger following of support due to the simplicity of the message and ease at which it could be shared across social media. Highlighting images of the young, vulnerable women who had been kidnapped evoked an instant emotional cognitive opening with wider audiences and the crucial choice by the activists to optimize a slogan in English combined with the language choice of "our" girls meant that wider global audiences could feel part of the solution. The "our" was ambiguous so that the global community could feel as if they were part of the protectorate group urging governments to act. The core group of activists launched a website making 13 clear demands of the Nigerian government, which includes (1) demanding that no resource is spared in attempts to return the Chibok girls, (2) the government put forward strategies to better ensure incidence like this do not happen again, and (3) the government instate a better reintegration system for returned victims of kidnapping. In this case, the website notes that the activists have developed a Verification, Authentication and Reunification System (VARS) for abducted citizens who are rescued. VARS is a holistic approach for rehabilitation, reintegration, and resettlement of victims of abduction (Bring Back Our Girls n.d.).

Measuring success can be difficult in cases where the effort needs to take place downstream, between governments confronting terrorist activities. This grassroots counterspeech campaign has had some notable successes. Due to the campaign, the United States and several European governments mobilized resources to assist in the search. The Nigerian government carried out several operations against Boko Haram and currently of the 276 girls that were abducted, the official Bring Back Our Girls website notes that 57 girls were able to escape, 107 were released through negotiations or government intervention, and 112 are still missing. While the hashtag drew global and measurable attention and resources to help the Chibok girls, that attention also made them an asset to Boko Haram, meaning that the group felt emboldened to request significant ransom payments for prisoner releases or capitalized on the girls as their own metric of success (Parkinson and Hinshaw 2021b).

When national government actions depend on peer pressure from international attention, it is hard to keep the level of pressure needed for sustained focus. It is also difficult to measure how much government resources have gone toward these efforts when government transparency is lacking. The Bring Back Our Girls movement continues, and in the eight years since the abduction of the Chibok girls, a slow stream of attention has continued to ensure global awareness and pressure remains. A string of documentaries, books, and media articles chronicle the continued struggle to counterterrorism in Nigeria and across sub-Saharan Africa, forcing elected officials to address these issues and pose solutions. The activists continue to keep track of the return of Chibok girls to remind the public where victories take place and where the government's lack of certain actions continues.

In both the Black Lives Matter and the Bring Back Our Girl case studies, grassroots efforts were able to galvanise wider public and international support

to force certain strategic sociopolitical changes to counter different forms of violent extremism. Both cases give counterspeech practitioners some clear takeaways. When a campaign has clear goals and calls to action, measurement and evaluation against those goals is feasible. Simple slogans open the door to mass appeal. The use of social media across platforms combined with offline engagement moments creates a self-feeding loop of campaign content and the ability to reach wider target audiences to join in the activism. When this is paired with clear campaign messages that humanize victims, campaigns yield higher emotional impact, acting as a cognitive opening to target audiences. In both of these cases, the main goal was not necessarily to change the minds of extremist members or those already sympathizing with hate-based ideologies. For that form of targeted counterspeech, a higher degree of practitioner expertise and public-private partnerships are often needed.

Public-Private Partnerships

Partnerships between social media platforms and NGOs create an interesting counterpoint to organic counterspeech. Public-private partnerships allow for a greater combination of grassroots activism with higher precision in online targeting, data collection, and evaluation. Campaigns to unite the public or champion a cause are meant to bring the highest number of people together. Campaigns that are meant to reach low-prevalence, high-risk audiences of either vulnerable or extremist groups require a high level of oversight and precision. The following case studies review a partnership between YouTube and UK-based Muslim community activists challenging the so-called Islamic State (ISIS) in a campaign called Abdullah X and a Facebook Search partnership with a grassroots NGO combatting white supremacy in the United States.

Case Study 4: Facebook Search Redirect Partnership with Life After Hate

In March 2019, Facebook partnered with an NGO based in the United States called Life After Hate to launch a counterspeech initiative to redirect online users looking for white supremacy or neo-Nazi-related groups online to real-world support structures that could help them. Based on a series of research issues Facebook conducted and published on, the issue raised by many PVE/CVE practitioners was that there was no existing internal proactive mechanism by platforms to challenge or redirect processes of radicalization toward existing localized counterspeech content or resources. In other words, platforms have swathes of internal indicators, and NGOs make credible counterspeech and sometimes receive funding from tech companies, but the two processes remain largely separate (Saltman, Kooti, and Vockery 2021). The aim of the Redirect Methodology pilot (Moonshot 2023) was to test the capacity of online partnerships between social media companies and disengagement NGOs to turn passive online content searches with extremist

indicators into active engagement between an at-risk audience and a trained outreach professional off the platform.

To target more downstream, vulnerable, and increasingly extreme audiences, it is crucial that the NGO interfacing with the potential extremists has credibility with that audience and protocols for safety. Life After Hate is an organization that provides crisis intervention, education, support groups, and outreach to individuals trying to disengage from and leave white supremacy and violent extremist groups on the far-right in the United States. The organization manages the ExitUSA program for rehabilitation and was founded by former extremists, specifically for “helping people leave the violent far-right to connect with humanity and lead compassionate lives” (Life After Hate n.d.). If passive viewing and engagement with online extremist content can lead to active engagements extremist groups, this methodology attempted to replicate the model of passive searches leading to the opportunity for more active engagement with a localized online community that could facilitate disengagement from an extremist ideology.

Facebook was not the first to conceptualize a “Redirect Method”. It was originally created with search engine platforms in mind. Google Search, YouTube, and Bing have all deployed similar pilots; however, the model was reconfigured for Facebook’s search functionality since searches on Facebook for on-platform content are inclined more toward searching for individuals or groups, as opposed to wider search queries. Instead of introducing an individual to a playlist of content, this approach introduced an individual to an organization providing tangible support to leave a hate-based movement.

As such, measurement and evaluation were based on two metrics: (1) looking at Facebook’s metrics around click-through rates (CTRs) from the NGO campaign on Facebook search results to the NGOs website and (2) feedback from the NGO on increased web traffic and actual outreach from extremist group members looking for support. In the three months after the launch of the Life After Hate redirection, Facebook saw that when terms were searched related to the redirection list there was on average a 4% CTR. The average CTR is roughly 1.91% for search-related ads surfacing, with a healthy or successful CTR being between 4% and 5% (LOCALiQ n.d.). As reported by the NGO in the published study, the NGO website went from nearly zero referrals based on Facebook resources and CTR to an average of ± 200 new users to their websites per month. This period also correlated with a few dozen new longer-term engagements with individuals looking for assistance in leaving white supremacist organisations and movements. This reflects, anecdotally, a redirect initiative taking a passive search function and turning it into active counterspeech engagement with those furthest down the process of radicalization.

While this project highlights the potential for online counterspeech to lead to real-world impactful evaluation metrics to counter extremism, it is limited in that it takes concerted efforts and trust between a private company and an often underresourced NGO. It is also hard to find global NGOs that have the

nanced credibility and infrastructure to carry out disengagement and rehabilitation work with vulnerable extremist group members. Since this pilot, Facebook has expanded these partnerships to include Redirect programs in Australia, Germany, Indonesia, and the UK (Facebook n.d.).

Case Study 5: YouTube Partnership with Abdullah X

In early 2014, the threat from ISIS propaganda, recruitment, and brutality was visceral and all documented and exploited online. Western governments, think tanks, and experts alike were facing an uphill task in curtailing the online proliferation and allure of ISIS. One large concern was the exponential rise in young Muslim males and females being attracted to the well-crafted and “lifestyle-based” communications created by ISIS, resulting in unprecedented numbers of people leaving the relative sanctity of their Western lives to go and join the “struggle” for an “Islamic State” (Institute for Strategic Dialogue 2015). Out of this emergency and state of flux emerged a campaign that aimed to alter the scope and scale of what “counternarratives” or counterspeech within CVE efforts could achieve.

“Abdullah-X” (hereafter AX) was created in 2012 by grassroots activists from within the British Muslim community to offer both critique and counter messaging to Jihadist narratives through the medium of animation and thought-provoking dialogue (Institute for Strategic Dialogue 2017–). Using an avatar-based character depicting a streetwise teenager from East London, AX homed in on both popular culture and subcultural references to address issues such as extremist group conspiracy theories, social inequality, racism, and generational grievances. These were packaged in often grainy and gritty short, animated cartoons, with the narrative arcs weaving in difficult social issues through broader themes of identity, belonging, loyalty, and duty. Rather than trying to reach a broad cohort of young people in prevention efforts, this campaign series was aimed at young British Muslims who were already questioning or sympathizing with extremist narratives, as a low prevalence but high-risk audience. The theory behind the campaign was that most young people found allure by Islamist narratives because the extremist groups were addressing real concerns that no other outlets – such as within school, through parents, or within mosque – were addressing. Despite a well-thought-out idea from a creative and communications angle, AX could not reach the audiences it needed since online targeting to smaller audience segments can be challenging.

Working with the Radicalization Awareness Network (RAN), the AX project was selected for a pilot campaign in partnership with YouTube, which would provide “back-engine” support to drive the content to specific target audiences. Six short videos were launched on social channels set up to disseminate content and drive engagement. AX was analyzed as successful largely due to the first-person perspective, merging authenticity with relatable tones of defiance while addressing social realities (Braddock 2020). The project had a cross-platform strategy with presence across a YouTube channel, Twitter, and Facebook

profiles as well as a stand-alone website to counter the increasing prevalence of extremist content on such platforms. Using targeted marketing techniques, authentic feeling content, and an engaging visual style, the campaign videos reached over 52,000 viewers (70% from ads) and over 1 million “impressions” in the UK over a six-week trial period fostered by the direct support provided by YouTube. By addressing contemporary and controversial issues relevant to young Muslims, the channel was assessed as having impressive engagement from some of the hardest to reach at-risk audiences compared to other counterspeech content online (RAN 2015). The campaign even prompted a counter response directly from ISIS.

The use of animation, tech tools, credible messaging, and powerful networking through public–private partnerships enabled AX to push the boundaries of what counterspeech could do from both a campaign context and a creative context (RAN 2019). What made the project “work” was its ability to be seen, scaled, and packaged effectively, with a relevant message and an accessible as well as familiar medium (animation). Similar to the previously mentioned initiative, there are difficulties in scaling success models of this kind because they rely acutely on partnerships between grassroots activists and large private tech companies. It can be difficult to sustain funding relationships as priorities for companies shift.

Measurement and Evaluation

As the five case studies discussed in this chapter highlight, counterspeech can take on many forms and be evaluated in a variety of ways. When it comes to online campaigns, platform-specific and cross-platform tools can help practitioners and researchers track and consolidate measurement of reach and engagement in near real time, facilitating the oversight of efforts to evolve with the threat or understand quickly whether a campaign is, or is not, reaching its intended effect. Most larger platforms have comprehensive ads marketing centers with targeting tools that can easily be reappropriated for practitioners engaged in PVE/CVE counterspeech. Facebook’s Counterspeech Site lists a range of resources, including guidance for non-profits for ads targeting strategies with tools and products to leverage (Facebook n.d.). YouTube also houses a program for assisting nonprofit organizations to connect with supporters, volunteers, and donors (YouTube n.d.). It might not always be intuitive for activists and practitioners to engage with ads targeting or wider marketing tools online, but even bringing in a younger researcher or intern with better digital fluency can bring innovation to the measurement and evaluation space.

When seeking to define success parameters, much is dependent on how the campaign’s outcomes are devised from its original objectives (Zamir 2022). Such dependencies also affect and shape associated elements such as advertising budgets, the link between effectiveness and efficiency, to choose or overlook capturing sentiment alongside data, and whether intended effects offered good

value once unintended effects were also factored in. A key issue within this is the use of paid reach and its opposite option of seeking organic reach and presence (Mina 2017). What makes counterspeech “good” counterspeech?

The ability to define and know the target audience, as previously discussed, is a necessary first step. This means that the most basic success criteria should be the confirmation that the counterspeech content has reached its intended audience. Reach statistics try to suggest that the content has been “served” to the people it was intended to be served to. Ads marketing tools can be used to target audiences that often include criteria around broad age, gender, and location, while content-specific data should usually show the duration of engagement, likes, shares, and retention statistics. These “reach” metrics help build a richer picture of who the content is attracting engagement from. Reach metrics alone do not allow a practitioner to know whether the content has had an impact on the intended audience. Reach by itself is a vanity metric. However, added to this mixture are factors such as the size and scope of the campaign, duration, and expectations placed upon the campaign from its benefactors, creators, or stakeholders (GIFCT 2021). In essence, success looks at “how many”, “how few”, or “how often” the content is engaged with, consumed, or reacted to. Success is dependent on how you the creator defines their initial parameters and call to action (Carlson 2021). If the evaluation can show that in the first instance the counterspeech campaign has reached the intended target audience and that the audience engaged with or acted upon that content in the way the practitioner had intended (a successful “call to action”), then impact can be measured.

Evaluation Methods and Activating Audiences

The more innovative and successful campaigns, whether holistic in nature or through public–private partnerships, have homed in on combining qualitative and quantitative metrics for understanding and building upon successes. At the heart of this issue lies the methodological battle of the objective (to counter a hate-based ideology) to review the statistical power of quantitative measurement and the interpretive discourse analysis elements of qualitative methods, which is discussed in detail by Garland and Buerger (chapter 3) in this volume (see also Braddock 2020; Jigsaw 2022). Capturing sentiment within a counterspeech context is essential if campaign efforts are to understand what type of knowledge, awareness, and behavioral change has or has not occurred, alongside how these elements have contributed to an overall “effect” of the campaign’s presence in the domain of its intended target audiences. The emergence and reliance on “big data” and reach metrics have given some counterspeech analysis a somewhat skewed understanding of outcomes, sometimes equating pure reach and impressive statistics with actual behavior change. However, when it comes to evidencing this behavioral change, numbers do not add up to tangible connections between reach and “change” (Zuboff 2019; Schick 2020).

One reason for this anomalous relationship can be found in how campaigns understand or rationalise mass communications practices against niche or hypernuanced needs of sectors like PVE/CVE or countering disinformation. This is where the skills of audience segmentation, layered messaging, and A-B testing are of utility. If a campaign is seeking behavioral change and is within the paradigm of counterspeech, knowing whose behavior needs to change is essential, but equally important would be to know what this change should look like in relation to adversaries, supporters, and observers of the phenomenon in question.

Strategic communication methodologies provide potential signposts for elements of success criteria. Looking firstly at content, however it is conveyed, content is what audiences see, hear, consume, and experience. Static content that is ostensibly a one-way vehicle for passive consumption is unable to inspire active engagement. Audiences are now accustomed to interactive and immersive content, which in some cases extends to interoperability of certain content forms to be repurposed, reused, and reworked through audience action. User-generated content has become a norm, and therefore strategic communications-based efforts can push through static messaging into the realm of audience experience, co-creation, and participation. Extremist ideologies allow a certain “slack” or tolerance to be repurposed according to its climate and context (Francis 2014). Counterspeech content can also consider what role it plays in helping audiences to actively shape their next steps as opposed to just seeking to direct them. The future of this work depends on its ability to “activate” audiences through value propositions, sentiment alignment that drives choice, and preference needs as well as incentivization of audiences to be part and parcel of social change through individual change.

The traditional “battle of ideas” is now subsumed within broader skirmishes of influence, taste, and the “incentivization of self”. Audiences want and need outcomes that sit beyond notions of moderate norms and branch off into realms of benefit calculations, gratification needs, and being “on-trend”, even regarding extremism. Tech tools and greater access to these tools can encourage practitioners, creatives, and activists alike to innovate and work in more nuanced ways to attract and retain audiences. Active engagement is about communication being able to effectively harness the right “touch points” that drive people to take on ideas, sentiments, and processes that would ultimately align with the objectives of counter speech and CVE.

The Future of Counterspeech

Counterspeech must evolve with the times to be effective. The rate and pace of information exchange, changes in audience dynamics, and the rise of technology-centric engagement have emboldened diverse actors to add their voices, narratives, and ideas to a social world that is already fragmented. Counterspeech content must also consider what role it can play in helping audiences to

actively shape their next steps as opposed to just seeking to (re)direct them. The future of this work depends on the ability of counterspeech to “activate” audiences through value propositions, sentiment alignment that drives choice, and incentivization of audiences to be part of positive social change. From the perspective of understanding how to engage audiences, traditional knowledge and tactics regarding target groups as relatively passive recipients have made way for a new hybrid audience dynamics. To some extent, even the notion of “the audience” has been repackaged into a more fluid understanding of audiences now becoming and behaving more like active agents. Interest groups create their own online speech, acting as both gatekeepers and gate crashers depending on need or situational reality. It can be argued thus that the relationship between narratives and audiences has been repurposed alongside the “retrofitting” of different value systems on top of existing structures associated with mainstream social life.

The need to engage audiences in a credible, sensitive, and activated way remains at odds with the immense scale and speed of online activity, which can lead a well-crafted campaign to be lost in the fray of wider online noise without algorithmic optimization. To date, some testing has emerged around the automation of counterspeech and usage of artificial intelligence to create and dispense counterspeech to online audiences. This has largely been tested in limited, controlled online capacities to tackle hate speech.³ While AI campaign creation systems have frequently proven problematic and controversial when launched in “the wild” of open online ecosystems, there is potential for these tests to advance and tackle the issue of scale and speed. As discussed in Case Studies 4 and 5, hybrid models leaning into existing algorithmic support systems combining partnerships between tech companies and PVE/CVE practitioners can yield both innovative and effective counterspeech approaches.

Counterspeech must also learn from other campaigns and PVE/CVE efforts. To date, the vast majority of counterspeech has been launched in one of three ways: (1) through organic networks, (2) via ads targeting, and (3) through search redirection. All three methods have progressed in the last ten years, contributing to counter hate and violent extremism online. However, the measurement and evaluation of these approaches is largely based on reach and engagement, often overlooking behavioral or sentiment change analysis. In the aftermath of the Snowden Revelations and scandals like Cambridge Analytica, governments have introduced increased user data privacy laws (e.g., ePrivacy Directive and General Data Protection Regulation [GDPR]). While this is overall positive for user protections, this has also led to restricted researcher access to data and the recall of certain tools, such as CrowdTangle, that assisted counterspeech practitioners in evaluation work. Although it remains harder to build nuanced evaluation frameworks, it is crucial in answering whether counterspeech has been a success. Direct partnerships with tech companies, as discussed in the case studies within this chapter, show how innovative measurement and evaluation tactics can be co-developed and should be encouraged.

Lastly, counterspeech must evolve to newer spaces online. Most counterspeech remains launched on the same three to four larger platforms, where NGOs and practitioners feel most comfortable. However, just a brief glance at the range of apps in use on your own phone – or your child’s phone – is a stark reminder that communication and engagement is increasingly cross-platform. From gaming-adjacent chat sites to more personalized chat forums to video streaming sites, there is a wider array of platforms that deserve more attention for counterspeech strategies. At the end of the day, as this chapter began, knowing the audience you want to reach is crucial and provides the information needed to understand which platforms need engagement. Go where your target audience spends the most time. Engage in the harder to reach spaces but not without foresight and planning around safety protocols, escalation pathways, and practitioner safety. The cat and mouse between PVE/CVE activists and violent extremist groups will continue to evolve. Diversified threats mean that refreshed tactics and innovation are needed to guide PVE/CVE work and counterspeech will remain a critical component, adding to and building from offline efforts to engage at-risk and extremist target audiences.

Notes

- 1 See also Bahador (chapter 4) in this volume, who discusses different counterspeech audiences.
- 2 See chapters 1 and 6 for further analyses of counterspeech on social media sites and gaming platforms, respectively.
- 3 See Guerini and Fondazione Bruno Kessler (2020) and counterspeech AI test portals such as “Automatic Generation of Counterspeech to Fight Hate Speech” (2022), which states “Please note: This system is a prototype and cannot be guaranteed to always generate appropriate responses. Any inappropriate responses expressed by the system should not be construed as reflective of the views or values of the researchers”.

References

- Aly, Anne, Stuart Macdonald, Lee Jarvis, and Thomas Chen. 2016. *Violent Extremism Online: New Perspectives on Terrorism and the Internet*. New York: Routledge.
- Ankel, Sophia. 2020. “Black Lives Matter: 16 Big Changes Since George Floyd Protests.” <https://www.businessinsider.com/13-concrete-changes-sparked-by-george-floyd-protests-so-far-2020-6>
- Anti-Defamation League. 2018. *Pyramid of Hate*. <https://www.adl.org>. Accessed 15 July 2022.
- “Automatic Generation of Counterspeech to Fight Hate Speech.” 2022. <https://huggingface.co/spaces/shaneweisz/AutoCounterspeech>
- Black Lives Matter. 2023. “About.” <https://blacklivesmatter.com/about>
- Braddock, Kurt. 2020. *Weaponized Words*. Cambridge: Cambridge University Press.
- Bring Back Our Girls. n.d. “112 #ChibokGirls Are Still Missing.” <https://bringbackourgirls.ng>
- Buchanan, Larry, Quoctrung Bui, and Jugal Patel. 2020. “Black Lives Matter May Be the Largest Movement in U.S. History.” *The New York Times*, 3 July. <https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html>

- Butler, Octavia, and Black Lives Matter Global Network Foundation. 2021. *Black Lives Matter 2020 Impact Report*. <https://blacklivesmatter.com/wp-content/uploads/2021/02/blm-2020-impact-report.pdf>
- Carey, James. 2009. *Communication as Culture*. 2nd ed. New York: Routledge.
- Carlson, Caitlin Ring. 2021. *Hate Speech*. Cambridge, MA: MIT Press.
- Cauberghs, Olivier. 2022. *Does Exposure to Online Extremist and Terrorist Content Influence the Mental Wellbeing of Practitioners?* University of St Andrews. 30 May. <https://www.linkedin.com/pulse/does-exposure-online-extremist-terrorist-content-mental-cauberghs/>
- Conway, Maura. 2021. "Online Extremism and Terrorism Research Ethics: Researcher Safety, Informed Consent, and the Need for Tailored Guidelines." *Terrorism and Political Violence* 33 (2): 367–380. doi:10.1080/09546553.2021.1880235
- Cram, Ian. 2019. *Extremism, Free Speech and Counter Terrorism Policy and Law*. Basingstoke, UK: Routledge.
- Dawson, L. 2019. "Clarifying the Explanatory Context for Developing Theories of Radicalization: Five Basic Considerations." *Journal for Deradicalization* 18: 146–184. <https://journals.sfu.ca/jd/index.php/jd/article/view/191>. Accessed 10 September 2023.
- Dekens, Nico. 2020. *Vicarious Trauma and OSINT: A Practical Guide*. <https://osintcurio.us/2020/06/08/vicarious-trauma-and-osint-a-practical-guide/>
- Donzelli, Silvia. 2021. "Countering Harmful Speech Online. (In)effective Strategies and the Duty to Counterspeak." *Phenomenology and Mind* 20: 76–87.
- Facebook. n.d. *The Redirect Initiative*. Counterspeech Site: Supporting the Voices That Are Engaged in Counterspeech. <https://counterspeech.fb.com/en/initiatives/redirect/>. Accessed 31 August 2022.
- Francis, Matthew. 2014. *How Do We Prevent Radicalization*. <https://www.theosthinktank.co.uk/comment/2014/09/26/how-do-we-prevent-radicalisation-by-matthew-francis>. Accessed 12 July 2022.
- Fukuyama, Francis. 2018. *Identity*. London: Profile Books.
- Gendron, Angela. 2016. "The Call to Jihad: Charismatic Leaders and the Internet." In *Violent Extremism Online: New Perspectives on Terrorism and the Internet*, edited by Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen, 25–44. New York: Routledge.
- Glazzard, Andrew. 2017. *Losing the Plot*. International Centre for Counter Terrorism Publications. <https://icct.nl/publication/losing-the-plot-narrative-counter-narrative-and-violent-extremism/>. Accessed 12 July 2022.
- Global Internet Forum to Counter Terrorism. n.d. *Resource Guide*. <https://gifct.org/resource-guide/>. Accessed 12 August 2022.
- Global Internet Forum to Counter Terrorism. 2021. *Content-Sharing Algorithms, Processes and Positive Interventions*. <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI2-2021.pdf>
- Guenther, Lars, Georg Ruhmann, Jenny Bischoff, Tessa Penzel, and Antonia Weber. 2020. "Strategic Framing and Social Media Engagement: Analyzing Memes Posted by the German Identitarian Movement on Facebook." *Social Media + Society* 6 (1): 1–13.
- Guerini, Marco, and Fondazione Bruno Kessler. 2020. "Generating Counter Narratives against Online Hate Speech: Data and Strategies." *Hatometer*. https://magazine.fb.eu/wp-content/uploads/2020/07/TENURE_FINAL_PRESENTATION.pdf
- Hedges, P. 2017. "Radicalisation: Examining a Concept, Its Use and Abuse." *Counter Terrorist Trends and Analyses* 9 (10), 12–18. <http://www.jstor.org/stable/26351560>
- Institute for Strategic Dialogue. 2015. *Becoming Mulan – Female Western Migrants to ISIS*. <https://www.isdglobal.org/isd-publications/becoming-mulan-female-western-migrants-to-isis/>. Accessed 12 July 2022.

- Institute for Strategic Dialogue. 2017–. “Campaign Toolkit.” <https://www.isdglobal.org/action-training/campaign-toolkit-innovation-hub/>
- Jigsaw. 2022. *Prebunking Anti-Vaccine Narratives*. <https://medium.com/jigsaw/prebunking-anti-vaccine-narratives-an-effective-alternative-to-debunking-individual-false-claims-78f0047a8b47>. Accessed 15 July 2022.
- Keen, Andrew. 2012. *Digital Vertigo*. London: Constable Publications.
- King, Peter. 2018. *Building Resilience for Terrorism for Terrorism Researchers*. <https://www.voxpol.eu/building-resilience-for-terrorism-researchers/>. Accessed 6 September 2023.
- Lebron, Christopher. 2017. *The Making of Black Lives Matter*. Oxford: Oxford University Press.
- Life After Hate. n.d. “Home Page.” <https://www.lifeafterhate.org/>. Accessed 31 August 2022.
- LOCALiQ. n.d. *Click-Through Rate (CTR): Understanding CTR for PPC [PPC U]*. <https://www.wordstream.com/click-through-rate>. Accessed 31 August 2022.
- Massanari, Adrienne L. 2018. “Rethinking Research Ethics, Power, and the Risk of Visibility in the Era of the ‘Alt-Right’ Gaze.” *Social Media and Society* 1–9. doi:10.1177/2056305118768302
- McIlwain, Charlton D. 2020. *Black Software – The Internet and Racial Justice*. New York: Oxford University Press.
- McIntyre, Lee. 2018. *Post-Truth*. Cambridge, MA: MIT Press.
- Mina, An Xiao. 2017. *Memes to Movements*. Chicago: Beacon Press.
- Moonshot. 2023. “The Redirect Method.” <https://moonshotteam.com/the-redirect-method/>. Accessed 6 September 2023.
- Parkinson, Joe, and Drew Hinshaw. 2021a. *Bring Back Our Girls: The Untold Story of the Global Search for Nigeria’s Missing Schoolgirls*. HarperCollins.
- Parkinson, Joe, and Drew Hinshaw. 2021b. “How the ‘Bring Back Our Girls’ Tweets Changed a War in Nigeria.” *The Wall Street Journal*, 20 February. <https://www.wsj.com/articles/how-the-bring-back-our-girls-tweets-changed-a-war-in-nigeria-11613797261>
- Radicalisation Awareness Network. 2015. “Counter Narratives and Alternative Narratives.” RAN Issue Paper. https://home-affairs.ec.europa.eu/system/files/en?file=2020-09/issue_paper_cn_oct2015_en.pdf. Accessed 6 September 2023.
- Radicalisation Awareness Network. 2019. *Collection of Practices*. EU Home Publications. https://home-affairs.ec.europa.eu/system/files/2021-05/ran_collection-approaches_and_practices_en.pdf. Accessed 15 July 2022.
- Ransby, Barbara. 2018. *Making All Black Lives Matter: Reimagining Freedom in the 21st Century*. Oakland: University of California Press.
- Reed, Alastair, and Haroro Ingram. 2018. *A Practical Guide to the First Rule of CT-CVE Messaging*. https://www.europol.europa.eu/cms/sites/default/files/documents/reed_ingram_a_practical_guide_to_the_first_rule_of_ctcve.pdf
- Reeve, Zoey. 2020. “Repeated and Extensive Exposure to Online Terrorist Content: Counter-Terrorism Internet Referral Unit Perceived Stresses and Strategies.” *Studies in Conflict & Terrorism* 46 (6): 888–912. doi:10.1080/1057610X.2020.1792726
- Saltman, Erin, and Anne Craanen. 2022. “Breakout Discussion on ‘Ethical Challenges in P/CVE Research: OSINT and Digital Spaces.’” Paper presented at the International CVE Research Conference 2022, 24 May, Granada, Spain. <https://hedayah.com/events/cve2022/>.
- Saltman, Erin, Moli Dow, and Kelsey Bjornsgaard. 2016. *Youth Innovation Labs: A Model for Preventing and Countering Violent Extremism*. The Institute for Strategic Dialogue. isdglobal.org/wp-content/uploads/2016/07/YouthCAN-Labs.pdf.

- Saltman, Erin, Farshad Kooti, and Karly Vockery. 2021. "New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis." *Studies in Conflict & Terrorism* 46 (9): 1547–1574. doi:10.1080/1057610X.2021.1888404
- Schick, Nina. 2020. *Deepfakes*. London: Monoray.
- Speckhard, Anne, Ardian Shajkovi, Claire Wooster, and Neima Izadi. 2018. *Engaging English Speaking Facebook Users in an Anti-ISIS Awareness Campaign*. <https://www.icsve.org/engaging-english-speaking-facebook-users-in-an-anti-isis-awareness-campaign/>.
- Stoor, Will. 2017. *Selfie*. London: Picador.
- Trottier, Jonathan. 2013. *Identity Problems in the Facebook Era*. New York: Routledge.
- Tuck, Henry, and Tanya Silverman. 2016. *The Counter-Narrative Handbook*. ISD Global. https://www.isdglobal.org/wp-content/uploads/2018/10/Counter-narrative-Handbook_1_web.pdf
- White, Kesa, Jacob Davey, and Galen Lamphere-Englund. 2022. *Good Practices, Tools, and Safety Measures for Researchers*. <https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-PI-GoodPractice-1.1.pdf>.
- Whittaker, Joe. 2019. *Building Secondary Source Databases on Violent Extremism: Reflections and Suggestions*. Washington, DC: RESOLVE Network. doi:10.37805/rve2019.4
- Winkler, Carol K, and Cori E Dauber, 2014. *Visual Propaganda and Extremism in the Online Environment*. Carlisle, PA: Strategic Studies Institute and U.S. Army War College Press. <https://press.armywarcollege.edu/monographs/946/>
- Winter, Charlie. 2019. *Researching Jihadist Propaganda: Access, Interpretation, and Trauma*. Washington, DC: RESOLVE Network. doi:10.37805/rve2019.1
- YouTube. n.d. "YouTube Non-Profit Programme Overview – YouTube Help." Google Help. <https://support.google.com/youtube/answer/3545195>. Accessed 31 August 2022.
- Zamir, Munir. 2022. *Active Strategic Communications: Measuring Impact and Audience Engagement*. Washington, DC: Global Internet Forum to Counter Terrorism.
- Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism*. London: Profile Books.

9

CONCLUSION

Marcus Tomalin

Counterspeech is certainly not a new phenomenon in human communication. On the contrary, it has presumably existed for as long as toxic speech has been expressed and, therefore, it dates back to time immemorial. Nonetheless, the need to understand the various ways in which victims or bystanders can respond effectively to hate speech has perhaps never been more important. And this need to understand in greater detail the range of possible responses to toxic utterances gives conspicuous centrality to the closely related task of determining precisely what constitutes “effective” or “successful” counterspeech in practice, whether it takes the form of carefully pre-planned targeted counter-messaging campaigns or spontaneous, organic, improvised responses. The obvious implication is that if counterspeech can indeed be “effective” or “successful”, then presumably poorly constructed and/or badly expressed counterspeech can be ineffective or unsuccessful – or perhaps even dangerously counterproductive. If a highly charged argument is taking place on social media and one of the participants uses a racist slur, maladroit subsequent attempts to challenge the ideology behind the slur could risk escalating the situation, which in turn could risk increasing the resulting harm.

The ubiquity of social media and other forms of instantaneous electronic communication in our digital societies has made it extremely easy for hateful language to be weaponized and circulated rapidly, and at an industrial scale, whether directed at individuals or at specific groups. Since many of the most toxic online communications are sent anonymously, identifying the perpetrators is often nontrivial. Consequently, it is crucial to understand in more detail the linguistic characteristics of counterspeech, especially if it is viewed as a communicative practice for invectivity management (as Sebastian Zollner suggests in chapter 1). By repeatedly establishing relations of opposition or contrast to hateful utterances at various linguistic levels, counterspeech makes invectivity

visible, and therefore it becomes possible to examine how conversational interactions prompted by hate speech develop in digital discourses. Needless to say, this extremely complex task can involve a wide range of linguistic tools and communicative strategies, and these can be employed in different ways depending on the specific situational and media-related context, as well as the previous interaction history of the participants. Indeed, as Laura Caponetto and Bianca Cepollaro have emphasized (in chapter 2), it is this remarkable variability that makes counterspeech such an intricate phenomenon, and that in turn is why analytical techniques derived from the philosophy of language are so important. Whether a given counterspeech strategy is effective or not in a given context will depend in part on things such as the linguistic form of the toxic utterance (e.g., whether it conveys the toxic message implicitly or explicitly), the linguistic form of the response, and both the speaker's and counterspeaker's social roles or demographic associations. Different forms of hate speech prompt different forms of counterspeech, and identifying recurrent patterns in the rich and sometimes bewildering diversity is a task of real contemporary relevance and importance.

While it is a nontrivial undertaking to accurately quantify the amount of hate speech that is in existence at any one time, there is broad agreement that hate speech poses grave dangers for the cohesion of a modern society. This is partly because it can violate basic human rights and, therefore, if allowed to circulate in an unconstrained manner, it can lead to isolated acts of physical violence or more sustained conflicts on much a wider scale. For these reasons, hate speech is usually recognised as an extreme form of intolerance that can contribute to hate crimes. As Babak Bahador has shown (in chapter 4), counterspeech can often play an important role by minimizing the risks associated with hate speech. It does this by influencing in very specific ways the audiences to which it is directed. Some of the effects counterspeech produces include reframing (i.e., using empathy to rehumanise groups targeted by hate), cognitive dissonance (i.e., making hateful speakers feel uncomfortable by highlighting the logical inconsistencies of their arguments/statements), and the spiral of silence (i.e., the pressure hateful speakers experience to disengage due to the fear of online social isolation). For many democracies, therefore, the difficult task is to find a way of legislating to curb the use and dissemination of hate speech while also safeguarding the citizens' right to express their opinions freely. To take a concrete example, Article 10 of the UK's Human Rights Act (1998) states explicitly that "everyone has the right to freedom of expression", and the Article subsequently clarifies that this includes "freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers". However, the very next paragraph adds that this freedom "may be subject to formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society" (UK Government 1998). In legislation such as this, the tension between these two positions is evident and unavoidable. While many countries have similarly moved to introduce laws that

place appropriate constraints on communications deemed unacceptable from a legal perspective, others (most conspicuously the United States) have been largely disinclined to introduce extensive hate speech legislation, believing that the resulting constraints would delimit freedom of expression in undesirable ways. This conspicuous diversity of legislative practice around hate speech reveals the complexity and ideological intricacy of the underlying task.

As Jacob Mchangama and Natalie Alkiviadou have argued (in chapter 5), the prevailing censorship-based methods of dealing with hate speech may not be fit for purpose in the modern world. Currently, a given social media platform will either remove hate speech in an unprompted manner (sometimes by initially deploying an automated hate speech detection system to identify problematical posts) or else they will do so retrospectively; that is, *after* a user has made a formal complaint and if the human assessors agree that the post contravenes the policy or guidelines that the corporation has explicitly adopted. But why should unelected corporations be the gatekeepers of free speech in modern digital democracies? What happens if those corporations benefit financially by allowing hate speech of certain kinds to circulate? But even if these legitimate concerns are disregarded, it can be argued that centralized censorship of this kind hinders the autonomous and organic ways in which different communities can seek to counter hate speech. Arguably, the current paternalistic approach of privatized censorship is contrary to the essence of freedom of expression, which includes the right to seek, receive, and impart information and ideas of all kinds. While reflecting upon this crucially important issue, it is useful to (re)consider precisely how censorship is manifest in practice by social media companies. Just because a significant percentage of hate speech is communicated on the platforms such companies provide, it would be naïve to assume that all such platforms censor hate speech in essentially the same way. For instance, the U.S.-based Center for Countering Digital Hate has shown that Twitter fails to act on 99% of the hate speech tweeted by Twitter Blue subscribers, suggesting that such users are covertly accorded greater freedom of expression than non-subscribing users. Seemingly, even free speech can be sold for financial gain. To take a contrasting example, the quarterly statistics released by Meta suggest that the amount of hate speech on Facebook and Instagram is currently falling (Center for Countering Digital Hate 2023). In the first quarter of 2023, for example, only 1 or 2 of every 10K content views on the two platforms contained hateful material (Meta 2023a). Meta generates these data by using the prevalence metric, which uses samples of content views. Specifically, the estimated number of views that show violating content is divided by the estimated number of total content views on those two platforms. Therefore, if the prevalence of hate speech were 0.20% to 0.25%, that would mean that for every 10K content views, 20 to 25 (on average) contained content that violated Meta's standards for hate speech (Meta 2022; see also Meta's policy on hate speech: Meta 2023b). These numbers imply that Meta is actively reducing the amount of hate speech that circulates on its two main platforms, and it does not seem

to be the case that certain users are accorded greater freedoms if they are willing to pay the company more money.

The intricacies of the issues involved here have recently been highlighted by the controversy surrounding Elon Musk's takeover of Twitter in October 2022. Reactions to Musk's acquisition of the company have been markedly mixed. Some have praised him for his planned reforms and for his calls for greater free speech, while others have criticized the changes he has introduced, pointing to a rise in misinformation, disinformation, harassment, and hate speech on the platform. As with all such complex matters, reliable quantitative data are hard to obtain. It is certainly true that since Musk acquired Twitter, specific controversial users who had been banned under the previous regime have had their accounts reinstated. This has happened to Andrew Anglin (founder of the neo-Nazi "Daily Stormer" website), to Liz Crokin (one of the biggest propagators of the QAnon conspiracy theory), and, perhaps most conspicuously, to former American president Donald Trump. However, reinstating controversial accounts does not necessarily lead to an automatic rise in hate speech. This is why interested independent groups and organizations have been trying to monitor the situation more closely. For instance, the Institute of Strategic Dialogue (ISD; Miller et al. 2023) has attempted to determine whether Musk's takeover had an impact on the amount of anti-Semitic hate speech in English that was circulating on Twitter. They did this using Beam, an automated hate speech detection system that ensembles 22 state-of-the-art machine-learning models in an attempt to classify a given tweet as being anti-Semitic or not. The system also uses five predetermined lexicons of hateful vocabulary. The results indicated that the volume of English-language tweets that were plausibly anti-Semitic had more than doubled in the period following Musk's takeover. More specifically, there were 325,739 English-language anti-Semitic tweets in the nine months from June 2022 to February 2023, with the weekly average number of such tweets increasing by 106% (from 6204 to 12,762), when compared to the period before Musk acquired the company (Miller et al. 2023).

The anti-Semitic tweets identified by the Beam system often made use of specific vocabulary to express the underlying ideological stance. For instance, one of the tweets presented in the ISD's analysis took the following form (Miller et al. 2023, 12):

THE KIKES AND HILTER MUST BE RELATED!!! THEY HAVE
EVERYTHING IN COMMON WHEN IT COMES TO
HURTING INNOCENT PEOPLE!!!!

"Kike" is a derogatory slur for a Jewish person and therefore this lexical item conveys the anti-Semitic tendency of the tweet, while the spurious comparison with Hitler is clearly designed to be polemical. Although the ISD study found that takedowns of such content had also increased during the same period, the rate of takedowns was not rapid enough to keep pace with the increase in anti-

Semitic tweets. Of course, given the subtleties of the human judgment that is involved in determining whether or not a given tweet has breached the Twitter community guidelines, it is sometimes a moot point whether such a message should be removed or not. However, even if the text were deemed sufficiently acceptable to remain visible on Twitter (as happened in this specific case), an appropriate form of counterspeech could effectively challenge the underlying ideology that has motivated this particular instance of anti-Semitism. In particular, the attempt to suggest similarities, or even connections, between the Nazis and the State of Israel could be usefully challenged and undermined.

While considering such examples, it is essential to remember that hate speech is not only confined to social media platforms. As Susana Costa, Bruno Mendes da Silva, and Mirian Tavares have shrewdly reminded us (in chapter 6), young people who regularly play online computer games are frequently exposed to cyberhate. One particularly concerning conclusion of their study is that there seems to be a direct link between exposure to hate speech and a greater tendency to practice it. In some ways this is not surprising, since most children learn through imitative practices, and many multiplayer video games provide collaborative learning scenarios. Consequently, their research provides a revealing case study of how hate speech can beget hate speech. The authors' emphasis on gamification and counterspeech as more effective ways than censorship of countering toxic messages recalls the work of Sander van der Linden (2023) on the need to inoculate users against fake news. In both cases there is a recognized need for greater education and instruction, to better enable users to cope as they become involved in potentially damaging online scenarios. Such proposals inevitably raise several practical issues: where and when should users be taught strategies for responding to hate speech (at home, at school, at university – at all of these?), and who should be tasked with providing the teaching? And even if we reach a position in the near future when all citizens of our digital societies are fully trained to use effective counterspeech strategies, we should remember that, in the modern world, a significant percentage of the (online) hate speech that circulates daily is either generated or distributed *automatically*. For instance, Joshua Uyheng and his colleagues showed that, during the COVID-19 pandemic, malicious bots hugely amplified the levels of racist hate speech that was communicated on Twitter (Uyheng, Bellutta, and Carley 2022). If human users, especially marginalized or vulnerable ones, are expected to respond successfully to this constant torrent of hate speech themselves, it is highly likely that their mental health will be adversely affected. It is now widely acknowledged that the (human) commercial content moderators who assess posts on social media platforms, to determine whether or not they contravene community guidelines, commonly endure levels of workplace trauma similar to those experienced by first responders (Steiger et al. 2021). They risk serious harm to their well-being as they are routinely exposed to posts containing hate speech, trolling, and/or graphic content depicting sexual abuse, child abuse, and other violent or distressing acts. Prolonged and repeated

exposure to such material can be damaging, and some moderators develop vicarious trauma, with symptoms that include insomnia, anxiety, depression, panic attacks, and posttraumatic stress disorder. The nature of their work means that they are unable to avoid traumatic content: they need to view the content to assess it, while seeking to meet demanding accuracy and speed targets. While it is true that moderators are an extreme example, since they encounter such large amounts of disturbing content on a daily basis, the problems they experience indicate the extent to which harmful online materials can have a deleterious impact on those who seek to engage with such materials directly.

Since hate speech is becoming increasingly automated, and since the psychological impact of repeated encounters with online hate speech is now better understood, the scenario raises the intriguing question of whether it is possible to *automate* counterspeech responses. Accordingly, the chapter by Marcus Tomalin, Shane Weisz, and James Roy (chapter 7) provides an overview of the current state-of-the-art systems that attempt to do precisely this. Given the speed at which such technologies are currently developing, it is certainly the case that this chapter will date far more quickly than any of the others in this book. Nonetheless, its importance will hopefully remain undiminished for some time to come simply because the kind of high-level experimental infrastructure it summarizes (e.g., training data, models, evaluation metrics) will remain essential for the foreseeable future, even as the specific details of these components continue to change every few months. For instance, it is not always immediately apparent to human users of social media whether a given response to hate speech is powerful and effective or not; therefore, any attempts to quantify the extent to which a particular reply constitutes “good” counterspeech necessarily require careful consideration of the syntactic, semantic, and pragmatic characteristics of the utterance (to name just three of the most salient properties). The chapter also seeks to facilitate future interactions between those who design language-based technologies and those members of non-governmental organizations who are seeking actively to tackle the problem of hate speech. If the process of designing and training automated counterspeech generation systems is profoundly informed by those individuals who are routinely required to write counterspeech, the resulting automated systems are likely to be much more nuanced and powerful. And such developments are particularly timely and relevant in the age of publicly available dialogue systems and large language models. State-of-the-art systems such as Blenderbot3 and GPT-4 have been specifically designed to engage in human-like conversational interactions with users, and since it is inevitable that they will encounter hate speech as inputs from time to time, they have been purposefully fine-tuned to deal with hate speech, in different ways. Indeed, it is remarkable just how different the counterspeech strategies adopted by such systems are. For instance, BlenderBot does not engage with hate speech and usually seeks to change the subject in a manner that is often awkward and problematical since it leaves the hate speech unchallenged. By contrast, ChatGPT-4 usually attempts to undermine the

ideological assumptions underlying the hateful language it has received as its input. The contrasting approaches adopted by the teams who developed these systems indicate that, at present, there is no consensus concerning the best way of automating counterspeech. Consequently, despite the impressive performance level of such systems when they are assessed in terms of general dialogue responses, the task of fine-tuning them to produce more effective responses to toxic inputs is a relatively recent research topic that urgently requires much greater exploration.

Motivated by important recent developments such as these, this book has attempted to provide a state-of-the-art overview of counterspeech from a range of disciplinary perspectives. While it is undoubtedly true that subgroups of researchers in fields as diverse as linguistics, law, sociology, philosophy, computer science, and information engineering now recognize the crucial importance of understanding what constitutes effective counterspeech, it is also true that *genuinely* interdisciplinary collaboration around this crucial topic has only just started to occur. All too often, academic research is siloed into narrow disciplinary (or even subdisciplinary) ruts. Different university faculties and departments have different administrative infrastructures, and the researchers associated with them usually attend different conferences and publish in different journals. A superb article about counterspeech that would certainly be accepted by *EJP Data Science* would be unlikely to be considered seriously by *Philosophy Compass*, even though both journals publish articles about counterspeech. Consequently, the existing conventions around academic specialisation tend to disincentivize interdisciplinary collaboration. Yet without interactions that bring together contrasting techniques and approaches from the arts, sciences, social sciences, and the humanities, it is hard to believe that a subject as complex and as challenging as counterspeech will ever be understood in sufficient detail. As Joshua Garland and Catherine Buerger argue so convincingly in chapter 3 of this book, sociolinguists and computer scientists are inclined to analyze the effective use of counterspeech in very different ways. The former may rely on detailed questionnaires and interviews focused on specific case studies, while the latter may use quantitative techniques that reveal hidden patterns in vast amounts of data. Yet *both* approaches are required if we want to understand in detail how humans deploy counterspeech most powerfully. Consequently, traditionally distinct groups of researchers will need to create opportunities to share their ideas more often – and this book has attempted to provide a platform for numerous researchers from many different disciplines who are keen both to share their own approaches to the topic while also learning more about the alternative perspectives their colleagues in different faculties and departments bring to the discussion.

It seems certain, therefore, that the future of research into counterspeech will be (must be!) increasingly interdisciplinary – and hopefully this important development will facilitate other distinctive research agendas that will undoubtedly emerge in the very near future. For example, despite the apparent

implications of the familiar phrase “hate *speech*”, such utterances frequently involve elements that are not overtly linguistic at all. For instance, when France lost to Argentina in the 2021 football World Cup, the black players in the team were subsequently subjected to a huge amount of racist abuse on social media. While some of the slurs took the traditional form of offensive phrases and sentences, many of them were expressed simply using sequences of icons and symbols, the most common being pictures of monkeys and bananas (Fitzpatrick 2021). Given the association between such images and long-established racist stereotypes, it is obvious that such communications should fall under the broad category of hate speech. Yet they require a different kind of counterspeech since they do not verbalize a specific ideology in an overt manner. There is no *explicit* argument that can be challenged and debunked. The communications are simpler and cruder than that. In a similar fashion, it is worth noting that the most widespread examples of toxic messages that are not exclusively linguistic are the hateful memes that circulate so widely online. These offensive messages generally use images in conjunction with words to convey a hateful communication. The prevalence of such posts on online platforms is so concerning that, in 2020, Facebook ran the Hateful Memes Challenge, in which participants were encouraged to develop new systems that could identify multimodal hate speech automatically (Meta 2020). Although initiatives like this are certainly welcome, they only partially recognize the true complexity of the problem, since these days hateful posts online are not only restricted to combinations of texts and images. On the contrary, they can (also) combine video, emojis, music, icons, and symbols to communicate offensive messages. In the analytical framework offered by social semiotics, “modes” such as writing, speech, image, music, and gesture are socially and culturally determined semiotic resources that can be deployed to create meanings in discourses. The various modes all have different affordances; that is, particular potentialities and constraints that impact the making of signs in specific representations (Kress and van Leeuwen 2001). Consequently, hateful memes can be classified as multimodal ensembles that bring together certain modes, not in a random manner but with the aim of forming a collective and interrelated meaning.

Since it has long been recognized that hateful communications frequently take the form of multimodal ensembles (especially when they appear on online platforms), it is intriguing that most academic research into counterspeech continues to focus predominantly on *text*-based responses (as evidenced by the examples considered in the chapters of this book). While such examples provide a perfectly sensible starting point, there is no theoretical or practical reason why carefully constructed counterspeech should not also take full advantage of the communicative power multimodality offers. If combinations of texts, images, music, icons, symbols, and the like enable hate speech to be conveyed in ways that are sometimes subtle, complex, and effective, then it is worth considering how counterspeech can utilize the full potency of the affordances associated with several modes simultaneously. And multimodal ensembles offer a much wider range of

semiotic options than utterances that are expressed using text alone. For instance, one possible response to a crude racist tweet (such as the ones sent to the French footballers in 2021) would be to modify and upload a popular meme template often referred to as “The Distracted Boyfriend” (see Figure 9.1).

The familiarity of this existing meme structure and the established culture centrality of the image lend argumentative force to the three noun phrases that are juxtaposed, with no overt linguistic coordination or subordination, to convey a specific ideological stance. Potentially, making use of existing formats such as this enables the counterspeech to articulate an extremely serious point but in a humorous manner that avoids any hint of preachy po-faced solemnity.

That said, there are of course risks in responding to multimodal hate speech with multimodal counterspeech such as that given in Figure 9.1. For instance, as an instance of counterspeech, the above meme could be criticized for using invective rather than logic. In other words, it simply insults all those who are attracted to racist social media posts, without explaining why it is undesirable to engage with such materials. And although the strategy adopted by the meme may produce some kind of (hopefully efficacious) cognitive dissonance in the person who posted the racist tweet, it could also simply serve to polarise the position of that individual, and possibly of those bystanders who were amused by the racist post and who share the sentiments it implied. In addition to these concerns, while exclusively text-based responses can be designed to eliminate any realistic possibility of plausible ambiguity, thereby mitigating the danger of misinterpretation, multimodal ensembles can sometimes introduce elements of uncertainty or nuance that lead to undesirable vagueness: if different people can interpret a given ensemble in slightly different ways, does it lose its capacity to



FIGURE 9.1 Multimodal counterspeech.

function as effective counterspeech? These are all issues that merit far more careful consideration than they have received to date, not least because automated dialogue systems and social media platforms will undoubtedly become ever more multimodal in the near future. And, as should be obvious from the preceding discussion, the task of analyzing multimodal ensembles is certainly one that benefits from genuinely interdisciplinary collaboration.

This conclusion has attempted, briefly, to summarize some of the central research themes that have been identified in this book, while also anticipating some of the developments in counterspeech research that are likely to emerge in the next few years. It is hoped that it is readily apparent that there are many interconnected topics that await more extensive consideration. And, obviously, this work must be undertaken with full awareness of developments in related academic fields which focus on topics such as misinformation, confirmation bias, and illusory truth. This is important because if counterspeech is seeking to change the worldview of haters and/or bystanders, there is necessarily an underlying assumption that, as a form of communication, it at least has the capacity to change some minds. If this were not the case, then it would be an almost entirely pointless enterprise. And any attempts to alter convictions or to challenge flawed ideologies in the modern world must be initiated and sustained with a deep understanding of how human rights can be preserved and secured in our digital societies that are increasingly interconnected by rapid communications technologies that enable so much hate speech to be posted anonymously on social media or else automated by artificial intelligence systems.

References

- Centre for Countering Digital Hate. 2023. "Twitter Fails to Act on 99% of Twitter Blue Accounts Tweeting Hate." <https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate>. Accessed 6 September 2023.
- Fitzpatrick, Michael. 2021. "French Police Probe Racist Abuse of Football Stars on Twitter." *Radio France Internationale*, 7 July. <https://www.rfi.fr/en/france/20210707-french-police-probe-racist-abuse-of-football-stars-on-twitter>. Accessed 6 September 2023.
- Kress, Gunther, and Theo van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.
- Meta. 2020. "Hateful Memes Challenge." <https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/>. Accessed 6 September 2023.
- Meta. 2022. "Prevalence Metric." <https://transparency.fb.com/en-gb/policies/improving-prevalence-metric>. Accessed 6 September 2023.
- Meta. 2023a. "Hate Speech." <https://transparency.fb.com/reports/community-standards-enforcement/hate-speech/facebook>. Accessed 6 September 2023.
- Meta. 2023b. "Hate Speech." <https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>. Accessed 6 September 2023.
- Miller, Carl, David Weir, Shaun Ring, Oliver Marsh, Chris Inskip, and Nestor Prieto. 2023. "Antisemitism on Twitter before and after Elon Musk's Acquisition." <https://beamdisinfo.org/wp-content/uploads/2023/03/Antisemitism-on-Twitter-Before-and-After-Elon-Musks-Acquisition.pdf>

- Steiger, Miriah, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–14. New York: Association for Computing Machinery. doi:10.1145/3411764.3445092
- UK Government. 1998. *UK Human Rights Act*. <https://www.legislation.gov.uk/ukpga/1998/42/contents>. Accessed 6 September 2023.
- Uyheng, Joshua, Daniele Bellutta, and Kathleen M Carley. 2022. "Bots Amplify and Redirect Hate Speech in Online Discourse about Racism during the COVID-19 Pandemic." *Social Media + Society* 8 (3). doi:10.1177/20563051221104749
- Van der Linden, Sander. 2023. *Foolproof: Why We Fall for Misinformation and the Way to Build Immunity*. London: Harper Collins.

INDEX

- #iamhere 7, 70–1, 78, 91
- Abdullah-X 10, 179, 185, 187
- Accuracy 74, 201
- Algorithm 3, 70, 72, 101–2, 107, 109, 173, 177, 180, 191; algorithmic amplification 87, 171
- Alternative for Germany / Alternative für Deutschland (AfD) 32, 45n4, 73
- Analysis and Response Toolkit for Trust (ARTT) 79, 80n17
- Artificial Intelligence (AI) 9, 102, 106, 108–11, 127, 159, 164, 180, 191, 192n3, 205
- Audience(s) 2, 8, 10, 28, 28–9, 32, 36, 52, 57, 59, 71–2, 77, 79, 80n5, 85–94, 96, 126, 135, 149, 172–80, 184–92, 192n1, 197
- Authoritarianism 101, 115–16
- Automated Speech Recognition (ASR) 152

- Black Lives Matter (BLM) 10, 176, 179, 181–84
- BlenderBot 159–60, 160, 201
- BLEU 152, 154, 161–2, 162
- Blocking 6–8, 17, 31, 54–8, 60–1, 110–11
- Bring Back our Girls 10, 179, 181, 183–4
- Bullying 114, 132; cyberbullying 29
- Bystander(s) 80n16, 88, 95–6, 148, 196, 204–5; bystander effect 77

- Causality 76–7; causal inference 68–9, 76
- Censorship 6, 8–9, 19, 51, 67–8, 100–4, 111, 115–16, 198, 200

- ChatGPT 10, 160, 160, 162, 162–3, 163, 166n8, 201
- Christchurch 171
- Civil liberties 50, 67
- Classifier 18, 73–4, 80n10, 150, 153–4, 159, 165, 165n3–5
- Cognitive dissonance 8, 85, 91–3, 197, 204
- Content moderation 5, 8, 103, 105–8, 110, 112–6, 175
- Contradiction 5, 18–9, 21–4, 26, 36, 37, 40, 45, 93, 127
- Corpus/corpora 6–7, 20–5, 33, 34, 37, 37–8, 73, 134, 151, 153–4, 159, 161; corpus linguistics 27, 32, 36, 39, 62n1
- Correlation 9, 74, 128, 131–4, 131–2, 152, 154–5, 166n6
- Council of Europe 110, 123, 134
- COunter NArratives through Nichesourcing (CONAN) 19, 45n2, 127, 150–1, 151, 161, 164
- Counterspeech taxonomy 5, 18, 127, 148, 150; affiliation 36; humor 177–8, 204; tone 19, 70, 72, 148, 156, 177–8, 187; visual communication 173, 188
- Counterspeech types 5
- COVID-19 pandemic 1, 100, 102, 200
- Crawling 9, 149–50
- Crowdsourcing 9, 150

- Dangerous Speech Project 2, 18, 147–8
- Decentralization 112, 115
- Democratization 101
- Devaluation 39

- Digital Services Act (DSA) 101, 106
 Disinformation 2–3, 5, 67–8, 73, 87, 100–6, 113, 115–6, 190, 199
 Diversity 155–6, 161–4, 180, 197
- Encoder-decoder 156–9, 158
 Ethnography 70–1, 77; digital ethnography 71–2
 European Commission 104, 106, 125
 European Union 101, 124, 128
 Evaluation 10, 19, 23, 35–6, 37, 39, 40, 50, 74, 79, 80n12, 93, 112, 114, 125, 129, 171–2, 179, 185–6, 188–9, 191; evaluation metrics 151–3, 155–7, 159, 163–4, 166n6, 201; Evaluation of Dialogue (EOD) 155
 ExitUSA 186
 Extension 20, 44, 157
 Extremism 10, 18, 21, 24, 87, 102–3, 105, 116, 171–6, 180–1, 185–6, 190–1
- Facebook 3, 10, 23–4, 69–71, 87, 103–6, 111–2, 114, 125, 139, 147, 159, 178–9, 182–3, 185–8, 198, 203; *see also* Meta
 Fact-checking 128
 Feedback 36, 152, 161–2, 164, 186
 Fine-tuning 10, 73, 153, 156–7, 159–61, 163–5, 201–2
 Fluency 10, 152–3, 156, 159, 164, 165n3, 188
 Framing 10, 70, 93, 157, 171–3
 Future of Free Speech (FFS) 100, 115, 116n1
- Gamification 9–10, 128, 134–5, 139–40, 200; video games 9, 88, 123–5, 127–35, 131–2, 139, 200
 Gegenrede 6, 19–20, 22–5, 25, 29–32, 45n4
 Generate Prune Select (GPS) 156
 Generative language model 151
 German Network Enforcement Act (NetzDG) 101, 103, 106
 Global Internet Forum to Counter Terrorism (GIFCT) 173, 180
 Godel 162–3, 162–3
 Gold-similarity 10, 153–4, 164
 Google 159–60, 186; GoogleBooks 21–2
- Hate group 8, 21, 85–9, 91–6, 130–3, 132
 Herding 8, 85, 91, 95–6,
 Human-in-the-loop 10, 151, 164
 Hypodermic needle model 96
- In-group/Out-group phenomenon 87–8, 93
 Intension 20
- Interactional 6–7, 17–20, 27, 31–2, 34, 41
 Interdisciplinary 67–8, 77–9, 202, 205
 International Covenant on Civil and Political Rights (ICCPR) 113–4, 116
 International Human Rights Law (IHRL) 9, 101–3, 106–9, 111–6
 Intervention(s) 91, 93, 180; positive interventions 172–3
 Interview method 71–2, 77–9, 80n6–9, 91, 134–5, 202
 Invectivity/invective 24, 26–9, 196
 Inverse Document Frequency (IDF) 154–5
 Islamic State (ISIS) 87, 171, 175, 185, 187–8
- Keyness 37, 38; *see also* keyword(s)
 Keyword(s) 36–7, 37–8, 150–1, 176; *see also* keyness
- Labeling 39, 74, 105
 Large language model (LLM) 157, 159–60, 164, 201
 Life After Hate 10, 179, 185–6
- Machine learning 18, 69, 72, 79, 115, 147, 152, 171, 199
 Macro analysis 75, *see also* micro analysis
 Messenger 71, 171–2, 174–5, 177
 Meta 21, 112, 148, 175, 198, 203; *see also* Facebook
 Metaphor 31, 50, 61, 139
 Micro analysis 19, 27, 32, 39, 41, 75; *see also* macro analysis
 Microsoft 125, 157, 159, 162
 Misinformation 5, 67, 104–5, 114, 171, 175, 178, 199, 205
- N-gram 20, 22, 154–5
 Narratives 25, 126, 129, 153, 174–5, 177, 181–2, 190–1; counternarratives 9, 19, 89, 102, 116, 126–7, 134–5, 139–40, 150–1, 173–4, 187
 Natural language processing (NLP) 69, 72, 79, 111, 157, 159, 165
 Negative collective stereotype(s) 93–4
 Neural network 151, 156, 164
 Nichesourcing 9, 19, 127, 150–1
- Otto vs. Paula 19, 32–41, 33, 34, 37–8, 40, 42–3, 45
- Persuasion 8, 19, 85, 91, 94–6
 Platform 171–7, 179–80, 185–8, 192, 192n2, 198–200, 202–3; cross-platform 172, 176, 180, 187–8, 192; multi-

- platform 171; social media platform 3, 5, 18, 89, 92, 94, 101–3, 106, 110, 113–5, 147, 175, 180, 185, 198, 200, 205; platformization 101; deplatforming 136, 138
- Predisposition 96, 125
- Presupposition 29, 41, 79; presupposition protest 17
- Pretrained 151, 153–4, 156–7, 159–62
- Prevent and counter violent extremism (PVE/CVE) 171–3, 175–6, 178–81, 185, 188, 190–2
- Qualitative 7–8, 36, 68–70, 72, 77–9, 80n11, 80n15, 90, 148, 189
- Quantitative 7–8, 22, 68–70, 72–3, 75–9, 131–3, 189, 199, 202
- Radicalization 105, 172, 185–7
- Rebuttal 22, 52
- Reconquista Germanica (RG) 7, 72–6, 78
- Reconquista Internet (RI) 7, 73–6, 78,
- Redirect method 185–6
- Reframing 8, 85, 91, 93–4, 197
- Regularized logistic regression function 73
- Regulation 68, 102–3, 105, 107, 110, 113–4, 116
- Relevance 28, 59, 153, 155–6, 164, 172, 197
- Renaissance Numérique 79
- Retrieval-based 156–7
- Rohingya genocide 3, 87, 102, 112
- Sequence-to-sequence 156 – 7
- Silencing 102, 110–1, 127, 139, 148; silencing effect 4; spiral of silence 8, 85, 91–2, 95–6, 197; silent crowd 71
- Sisi Ni Amani 89–90
- Stop Hate UK 151
- Targeting 100, 171–2, 174, 176, 179–80, 185, 188; ad targeting 176, 191; audience targeting 174, 180, 187
- Test data 161
- Toxicity 10, 50, 128, 131, 133–4, 152–4, 159, 161–2, 162, 164, 165n4–5,
- Training data 9, 18, 111, 149, 163, 165, 201
- Transformer 156–61, 158, 166n7; generative pretrained transformer (GPT) 151, 162
- Twitter 4, 7, 32–3, 34, 35–6, 38, 40, 44, 73–4, 90–1, 104–5, 125, 148, 175, 182–3, 187, 198–200
- Two-step flow 8, 85, 91, 94
- Unigram 155
- Word Error Rate (WER) 152
- Youth Civil Activist Network (YouthCAN) 10, 178
- YouTube 10, 45n4, 112, 125, 130, 135, 137, 139, 149, 175, 179, 185–8



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>