MAPPING AND ANALYSING HATE SPEECH ONLINE:
Opportunities and Challenges for Ethiopia
CONTENTS

Acknowledgements 4

Introduction and the scope of paper 5

Section 1: The concept of “hate speech” 9
   Defining hate speech 9
   International Legal Paradigms 10
   National legislation 11
   Hate speech and the Internet 12
   Hate speech vs. freedom of expression 13
   The targets of Hate Speech 14
   Hate speech in Divided Societies 15
   The intention of hate speech 16
   A question of balance 17

Section 2: A methodological toolkit for analysing hate speech 19
   Discourse analysis 20
   Content analysis 22
   Automated Techniques 27
   Towards a Framework for Further Analysis 28

Section 3: Hate speech in Ethiopia 30
   The uneasy relationship between the media and the state 30
   Online spaces, ethnic divisions, and the use of hate speech for political ends 33
   Towards a new research agenda 35

Appendix 1 – Definition of dangerous speech 39
ACKNOWLEDGEMENTS

We are grateful to the British Embassy in Addis Ababa for the support provided to envision and realize this project.

We are also grateful to the reviewers who have provided their precious advice and comments to the earlier drafts of this working paper.
Introduction and the scope of the paper

Hate speech lies in a complex nexus with freedom of expression, individual, group and minority rights, as well as concepts of dignity, liberty and equality, and is inherently context dependent. The definition of the term is therefore problematic, and often contested. Yet generally, hate speech refers to words of incitement and hatred against individuals based upon their identification with a certain social or demographic group. It may include, but is not limited to, speech that advocates, threatens, or encourages violent acts against a particular group, or expressions that foster a climate of prejudice and intolerance, which can lead to targeted attacks or persecution of that group through ‘othering processes’.

The issue of hate speech is complex – it serves as an illuminative lens for researchers and practitioners, governments and citizens, as well as perpetrators and victims, to consider some of the key paradigms that underpin models of statehood and democracy. The problem becomes all the more pressing and salient in deeply divided societies that are transitioning to democracy. In such contexts and in fragile environments, political entrepreneurs are able to politicise longstanding socio-economic inequalities that mobilise deeply entrenched feeling of injustice, domination and persecution towards certain demographic and social groups, with the ultimate objective of perpetrating systematic, targeted and widespread violence. Furthermore, as access to the internet changes the ways in which individuals and groups communicate, and produces new spaces for dialogue and exchange, it also brings with it risks and concerns over the ways in which these spaces may be instrumentalised for violent ends. For example, the use of hate speech across a variety of forums was well documented in the case of then 2007 Kenyan elections, and was considered as a key channel through which the violence was fuelled, which tragically claimed the lives of over a thousand people, and the displacement of over 600,000.¹

This working paper provides a framework through which hate speech can be identified and analysed. Given that hate speech is not just a concept that has been debated at the theoretical level, but one which has been deliberated by governments, legal practitioners, as well as civil society members – all of which have played a role in informing law and policy as well as projects and research – there are three key paradigms which are able to provide differing yet illuminative lenses to better understand the issue and the implications of hate speech acts.

- **Concept:** The practice of labelling hate speech as such can be particularly harmful in itself and serves to trump other fundamental rights, so caution must be applied when identifying hate speech acts as such. At what point does offensive speech become hate speech? Ultimately, there must be a boundary between the tolerable and the intolerable – in other words, under what circumstances does condemnation, insult or scorn be rightfully categorised as a message of persecution, inhumanity and degradation; and therefore unacceptable?
- **Context:** It is important to question the role which context plays in determining whether damaging and provocative speech amounts to hate speech. In the case of deeply divided societies, longstanding historical and social currents have often led to entrenched inequality and communal violence. Hate speech law is fundamentally shaped by ideological, historical, and social contexts, and is therefore specific to each country, whilst also sitting alongside international convention and legislation. Speech that attacks marginalised groups or incites violence must be rightfully analysed within its appropriate context to properly evaluate the dangerous implications it may have.

- **Transmission:** Which media should be taken into consideration when mapping hate speech? This is a particularly pertinent issue as access to the Internet increases, and social media platforms are being favoured over traditional outlets, such as radio and print media, for the expression of hate speech. Hate speech can also be disseminated in other unconventional arenas including SMS messages and song lyrics, and it is also important to consider the non-textual media through which hate speech may be carried, for instance through symbols and gestures, art and image.

Consideration of these three issues is able to inform methodological approaches that seek to identify, analyse and draw conclusions from the various instances of hate speech in different media. These approaches are diverse, and encompass evaluations of the content of the speech itself; an analysis of the discourse and therefore the social and political context underlying the speech act; as well as certain automated techniques that may be usefully employed, and are indeed facilitated by the changing nature of hate speech dissemination online. A critical evaluation of these approaches is therefore essential in order to understand the problems inherent to each method, as well as the trade off between the ability of providing a contextually grounded analysis of instances of hate speech and the possibility of processing large volumes of data. It also offers concrete case studies of projects which have innovatively analysed hate speech both online and offline.

The **aims** of this working paper are twofold.

**Firstly,** it is meant to provide an **introductory guide** for those interested in **mapping** and **analysing hate speech**, especially as communicated through online media and in divided societies. Section 1 offers an overview of the recent debates over the definition of hate speech, both from a scholarly and legal perspective. It briefly examines how different actors, from international bodies to national legislators and online companies, have sought to regulate and counteract hate speech, and stresses the importance of balancing competing principles to advance social inclusiveness as a “public good”. Section 2 surveys different methodological tools that can be adopted to map and analyse hate speech and the trade-offs between the ability of providing a contextually grounded analysis of instances of hate speech and the possibility of processing large volumes of data. It also offers concrete case studies of projects which have innovatively analysed hate speech both online and offline.
The second aim is to interrogate how a rigorous and academically informed analysis of hate speech can offer a novel terrain for engagement among different actors in divided societies, with a particular focus on Ethiopia. As it has been the case in other societies fractured across ethnic, political, and religious lines, at critical times such as before an election accusations of inciting hatred have been made by different actors to attack their opponents. The concept of hate speech advanced in these instances has tended to be boundless and to reflect more the interests and concerns of those referring to an act as hateful or able to incite hatred, rather than being based on an agreed definition. Section 3 reflects on the importance of bringing different actors around the same table, from members of the government and the opposition to bloggers supporting agendas at different ends of the political spectrum, and to engage in debates that can ultimately bound them to a shared definition of what is and what is not speech that can promote hate and violence, and to collectively recognize and develop measures to counteract it. A shared, academically rigorous and contextually grounded definition of hate speech and the mapping of instances that fall under it can help preventing governments and other political actors from politicising speech acts. Together with the analysis of how individuals are already opening spaces for dialogue on different platforms, it can offer a novel way to map where risks lie and where, on the contrary, people should be allowed to voice their opinions. In countries like Ethiopia, where calls for respecting and enforcing freedom of expression coming from external organizations, including human rights groups and donors, have tended to be ignored, this approach can open a new avenue for allowing dialogue among different forces in society, and with it a basis for a strong and vibrant democracy. The paper concludes by outlining the principles that should inform this new research agenda, aimed at mitigating ethnic, political and social fissures underlying debates on hate speech.
SECTION 1: THE CONCEPT OF HATE “SPEECH”
Hate speech lies in a complex nexus with freedom of expression, advocacy of hatred, and incitement to violence; and has to be understood in the historical, social and political context in which it occurs. It tends to be bold and confrontational, and may be communicated in a number of ways: via the media, on placards or graffiti, or increasingly on web chat-rooms and social media sites such as twitter and Facebook. The climate for hate speech is likely to become more conducive in situations where the political stakes are high, such as during elections or when undergoing regime change; or often when faltering economic growth, high levels of poverty and unemployment leads to hostility towards particular groups. Underlying the concept lie contested political terms such as equality, dignity, free expression, liberty and democracy, and as a result, the issue of definition becomes salient not only to describe and identify instances of defamatory and insulting speech, but in a more pragmatic sense to also legislate against such discourse at the national and international level. This section outlines some key conceptual issues associated with the term hate speech, focussing upon the tension between hate speech and freedom of expression; the intention of hate speech; the importance of the historical and social context it is located in; and the people who are the targets of this damaging practice.

Defining hate speech

Hate speech is an inherently complex and contested term. At a conceptual level, it is inflected by our understandings of two things: firstly, what constitutes “hate” and how we are to conceive hate against particular groups of individuals who find themselves the target of such speech; and secondly, how we are to view the potential consequences of speech and speech-related actions and their underlying intentions which are often multifaceted and deeply contradictory. Underpinning the concept of hate speech is the attempt and desire to defame and marginalise specific groups, often minorities. It plays a significant and instrumental role in the perpetuation of discrimination: not only is hate speech able to serve as a warning and threat to its victims, but also acts as a linguistic device to dehumanise and diminish targets, while simultaneously reinforcing a sense of identity and community among the perpetrators.

There is no universally agreed definition of what the term hate speech means – and consequently, there are many competing definitions of the term. It can be argued, however, that two general tendencies can be identified, which cut both through the policy and the academic debate. On one hand of the spectrum are definitions that are broad and comprehensive and aim at capturing occurrences of hate speech in their multi-faceted incarnations. Raphael Cohen-Almagor's definition of hate speech belongs to this first group.

“Hate speech is defined as bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic, and/or prejudicial attitudes towards those characteristics, which include gender, race, religion, ethnicity, colour,
national origin, disability or sexual orientation. Hate speech is intended to injure, dehumanize, harass, intimidate, degrade and victimize the targeted groups, and to foment insensitivity and brutality against them.2

On the other end of the spectrum are definitions which contest the concept of hate speech as too wide-ranging and therefore open to manipulation. Susan Benesch, for example, has challenged “hate speech” as too broad and ultimately problematic: it is found – in some form or guise – in almost all societies, including those where the risk of violence is limited; some hate speech in itself does not appreciably increase the risk of mass violence, even though it may cause significant emotional harm. She has advanced instead a narrower conception of “dangerous speech” to isolate acts that have a significant probability of “catalysing or amplifying violence by one group against another”.3 Benesch identifies five elements that should guide the identification of a dangerous speech act: i) the character and popularity of the speaker; ii) the emotional state of the audience; iii) the content of the speech act itself as a call to action; iv) the historical and social context in which the act occurs; and v) the means used to disseminate it (including the type of language adopted, since language can confer greater force on a speech act, e.g. if speech is in vernacular, a person from the area where that vernacular language is spoken may hear it differently than if it were in the national language). These five criteria (which are further detailed in Appendix 1 and discussed through the example of the UMATI project in Section 2) are aimed at focusing policy efforts on speech that has a special capacity to catalyse violence and reinforce the consequentialist stipulations embraced in international law, which places emphasis not on the expression per se, but on instead upon the danger it may cause, for instance in inciting violence and/or genocide.

In this section, both conceptual tendencies are explored, alongside legislative approaches. Together, these capture the challenges in employing particular definitions of hate speech, also indicating how they affect the methodologies to map hate speech and the instruments to regulate it and contain it.

**International legal paradigms**

Hate speech can have insidious consequences, ranging from emotional harm caused to the individuals and groups targeted, to the incitement of mass violence. The devastating corollaries of hate speech are evident in the examples of genocide or ‘ethnic cleansing’ in Rwanda, and the former Yugoslavia.4 It is therefore unsurprising that there have been pressures at the international level for efficacious legislation and creative responses to addressing and legislating hate speech and its consequences. The International Covenant on Civil and Political Rights (ICCPR) requires states to prohibit hate speech, which it considers to be:
“Any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility or violence”\(^5\)

The Council of Europe defines hate speech as covering:

“All forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.”\(^6\)

Definitions like these are broad and expansive; they cover both discrimination and violence, and need to be grounded in a specific context to understand the likely corollaries and dangerous implications that hate speech may have. Indeed, the European Court of Human Rights (ECtHR) stresses the importance of context in analyses and legislation of hate speech:

“It is only by a careful examination of the context in which the offending words appear that one can draw a meaningful distinction between shocking and offensive language which is protected by Article 10 [the free expression guarantee] and that which forfeits its right to tolerance in a democratic society.”\(^7\)

Without being rooted in the appropriate historical and social context, some approaches to defining hate speech can provide political actors with the jurisdiction to not only intervene in order to prevent individuals and groups from becoming the target of violent actions because of their ethnic, religious, or sexual identity, but crucially, they can offer justification for containing and silencing political contestation that is uttered in unconventional and potentially conflictual ways, but which nevertheless constitute part of a legitimate political opposition discourse or dissent. For this reason, international organizations have tried to suggest multi-layered approaches that – while seeking to identify elements potentially leading to human rights violations, crimes against humanity, and genocide – remain critically aware of the need to ensure freedom of expression. Adama Dieng, the UN Special Advisor on the Prevention of Genocide, for example, argued that the focus should be placed on the danger of hate speech, considering the severity, content, extent, imminence, likelihood and context of the speech.\(^8\)

### National legislation

As well as at an international level, hate speech is an issue that is addressed at the national level. Governments across the world have negotiated the balance between protecting freedom of expression which is necessary for the realisation of a democratic society, as well as preventing harm to individuals or minority groups as a result of hateful speech; be
this through the effects of the language itself or the violence it may incite. Governments have become acutely aware of the way in which it is able to fuel violence and contribute to tensions between groups. This is particularly salient in the context of deeply divided societies, where ethnic tensions and resulting inequalities may be politicised through the use of defamatory, pejorative and insulting speech. This was recently the case in Kenya, where drafters of the new constitution made explicit reference to the problem of hate speech as a result of its role in fuelling the 2007/8 post-electoral violence. Given the tensions between hate speech and freedom of expression, as well as its intersection in issues of human rights, equality and dignity, and laws governing the press, it is common for legislation regarding and applicable to hate speech to be found in various disparate pieces of legislation. It therefore becomes a subjective issue, and one which is reliant upon judicial interpretation, reflecting the importance of strong democratic institutions such as an independent judiciary.

Hate Speech and the Internet

The Internet has been celebrated for its ability to erode barriers between nations – even if states are increasingly fighting back and seek to re-territorialize the online space – but it has also created new walls separating new private spaces for expression. As of January 2014, Facebook had 1.24 billion users, who, if they were to be counted as citizens, would comprise the third – fastest growing – nation on the planet, after China and India. However, users of social networks like Facebook are subject not only to national law, but also to the terms of use set up by the companies owning the platforms where they choose to express themselves. While new online spaces allow for the realisation of many democratic paradigms through the provision of information and the facilitation of debate on political, social and economic issues, it also serves as a relatively safe space for the publication or posting of viewpoints which can be considered hateful and can lead to abuse and human rights violations because the internet affords a high degree of anonymity. Social media are also rapidly changing the nature of political engagement; extreme voices may use new platforms as they are often excluded from mainstream forums and traditional outlets. With expanding use and access to the internet, and specifically the proliferation of social media networks which offer spaces for the creation, publication, distribution and consumption of various types of media content, the problem of hate speech may be taking on new forms.

In this context, it is important to consider non-state/private sector responses to hate speech. Online service providers have advanced their own definitions of offensive speech, and regulate according to their Terms of Service.

Yahoo!’s Terms of Service prohibits the posting of “content that is unlawful, harmful, threatening, abusive, harassing, tortuous, defamatory, vulgar, obscene, libellous, invasive of another’s privacy, hateful, or racially, ethnically or otherwise objectionable”. 11

Youtube’s Community Guidelines makes explicit reference to hate speech:
Facebook’s terms are similar, forbidding content that is harmful, threatening and that has the tendency to stir hatred and incite violence

“We encourage free speech and defend everyone’s right to express unpopular points of view. But we do not permit hate speech: speech which attacks or demeans a group based on race or ethnic origin, religion, disability gender, age, veteran status and sexual orientation/gender identity.”

At the same time, as social media platforms are effectively self-regulating, operating through users’ reporting of language, images, videos, or audio that is perceived to harmful and then deciding whether or not the contested item should be removed, it is often the case that content that infringes the terms of use continues to be accessible until an individual or group of users or followers acts upon it. By this time, inflammatory messages can go viral and incite action in real spaces on the ground.

**Hate speech vs. freedom of expression**

In the opening address of the 83rd session of the Committee on the Elimination of Racial Discrimination (CERD), UN Deputy High Commissioner for Human Rights Flavia Pansieri asked:

“Where does the right of expression, which we all want to respect, stop and the need to sanction and prevent hate speech begin? What is the point in time when one right has to recognise that it cannot be exercised if it implies the violation of another one.”

Prima facie, freedom of expression can be seen as diametrically opposed to hate speech, since legislating against hate speech is typically conceived of as eradicating freedom of expression, and therefore eroding democratic gains. For a country to subscribe to a full and vibrant democracy, freedom of speech is vital to ensure that critical debates are held and a multiplicity of opinions voiced. In essence, freedom of speech is conducive to an effective “marketplace of ideas” and is therefore an essential means by which to foster a more vibrant and meaningful democracy, rather than simply elections and majority rule. To this end, unrestricted free expression may pave the way for enlightened and inclusive policy; the likelihood of authoritarian rule may be reduced; and it allows for citizens to better inform themselves of political, social and economic debates. Therefore, for the state to interfere and legislate against hate speech can be seen to curb diverse political expression and indeed embark upon a slippery slope of heightened intervention, which may ultimately lead to the suppression of political dissent and protest. On the one hand, it is important that speech which is able to form the basis of inter-ethnic cooperation and
understanding is not deterred; but on the other, it is important to avoid the corrosive impact that hate speech may have in alienating minority groups from society, thereby nourishing irredentist and secessionist movements.

However, Jeremy Waldron describes two ways in which hate speech can be considered a fundamental problem to society. He argues that societal inclusiveness should be conceived of as a ‘public good’ – it is something that societies should sponsor and be committed to. As people go about their daily lives, each group should be accepting of the fact that society is ‘not just for them; but it is for them too, along with all of others. And each person, each member of each group, should be able to go about his or her business, with the assurance that there will be no need to face hostility, violence, discrimination, or exclusion by others’. It is this sense of security that would be worth fostering in a society that values the equality and dignity of individuals. However, hate speech undermines this public good: ‘it does this not only by intimidating discrimination and violence, but by reawakening living nightmares of what society was like – or what other societies have been like – in the past.’ In the process, hate speech acts serve as a slow-acting poison to society, spreading incrementally, but effectively, making it difficult for the public good of inclusiveness to be fostered.

**The Targets of Hate Speech**

Hate speech is particularly pernicious and effective because of the way it constructs and politicises in-groups and out-groups. In this context, in-groups refer to the perpetrators of hate speech and the wider community that share their views and/or also partake in hate speech; out-groups refer to those groups that are the victims of the hate speech. In-groups are able to politicise particular social differences (e.g. race, ethnicity, religion, sexual orientation, gender) characterising an out-group in negative and dehumanized terms. In the process, the out-group is constructed as ‘other’ and effectively becomes distanced from salient social currents and prevailing opinions. The derogatory, pejorative and insulting slurs that characterise hate speech is particularly damaging because of its impact on vulnerable minorities, who have most likely in the past been hated or despised by others within the society. In the absence of hate speech, these vulnerable minorities are able to feel part of the body politic, as this absence stands as a tacit acknowledgement of their equal social status. However, hate speech compromises this ideal – it is a vituperative attack on the dignity of this group, both in their eyes, but also in the esteem of others. Its aim is to tarnish their reputation by associating phenotypical or other ascriptive characteristics like ethnicity, race, or religion with conduct or attributes that should defame and disqualify someone from being treated as a member of society in good standing.

In order to do this, in-groups may draw upon salient myths and histories about a particular historically oppressed social group, which still have purchase today. In appropriating this history, in-groups are able to capitalise upon previous historical experiences of out-groups, and in a particularly incendiary way, attempt to marginalise and frighten these groups. This has been seen in both Sudan and Rwanda. In the process, the in-group are able to not only crystallise its own identity, but also do so in positive terms. The issue of the targets of hate speech is therefore intimately connected to the wider social and political milieu in which it occurs.
Hate Speech in Divided Societies

In order for societies to be grounded upon the tenet of equality, and for the dignity of individuals to be afforded, it is important that we pay attention to the unique historical contexts specific to each country. As Ethiopian scholar Yared Mengistu argues, “we should not pretend that history has not affected the relative power among groups in a society where there has been a history of slavery, apartheid, holocaust, ethnic cleansing, or ethnic marginalization”.\(^{17}\) The impact and probable corollaries of hate speech are context dependent; they occur within a particular historical and social nexus, which has often led to the preferment of one group over another, and resulted in structural inequalities that persist today. When analysing hate speech, it is therefore important not to underestimate the impact of these historical and social currents.

This is particularly important in the context of deeply divided and democratising societies. Hate-fuelled discourse can incite violence because of the way it is able to frame historical events or create myths that propagate negative sentiment against particular groups or communities, and is therefore parasitic upon the historical and social context in which it occurs. Hate speech and mass violence are related in two ways: hate speech often precedes and fuels atrocity and is also an effective means – or conduit – to flare up hatred once incidents of mass violence have broken out. While they may be triggered by a specific event, they are usually the result of longstanding grievances that have accumulated over time, and have been fuelled by incendiary and offensive speech. Expressions fostering hatred based upon a group’s race or ethnicity can become so ingrained and entrenched in everyday vocabulary that the victimisation, marginalisation, persecution and ultimately massacre of out-groups is seen through a lens of normality. Such racial and ethnic affronts can foster the conditions for mass atrocity and violent acts against communities. Therefore, hate speech has an ability to create a climate of impunity, where the majority and institutions of the state become complacent in the face of minority oppression.

This was notably the case in the Rwandan genocide of 1994, where the use of powerful dehumanising metaphors to dehumanise the Tutsi ethnic group as “cockroaches” that needed to be exterminated, as well as the invocation of physical characteristics for example in the calls over radio broadcasts to “cut the tall trees”.\(^{18}\) It is important not to underestimate the role of the mass media in dissemination of hate speech, and therefore the incidence of violence and atrocity. Indeed, the UN Tribunal found that mass media hate speech can constitute “genocide, incitement to genocide, and crimes against humanity”.\(^{19}\) It is important not to underestimate the role of the mass media in dissemination of hate speech, and therefore the incidence of violence and atrocity. Indeed, the UN Tribunal found that mass media hate speech can constitute “genocide, incitement to genocide, and crimes against humanity”. This is important for any discussions of monitoring hate speech. This example is illustrative of the important point that an appreciation of historical and social context that hate speech occurs in is necessary to identify the coded language under which hides a vituperative attack upon a social group.
There is a delicate balance when it comes to negotiating between regulating hate speech and protecting minority groups and interests in richly diverse and deeply divided societies – indeed, it is difficult to regulate hate speech pertaining to ethnicity, when ethnicity itself is a prime vehicle for political assimilation and mobilisation. Indeed, hate speech can often be used as a tool to suppress opposition to the government (particularly one that is dominated by a particular ethnic group) and therefore perpetuate and entrench the subordination of historically oppressed minorities. As the state is not often a neutral arbiter of social relations, it is important that the rights of minority groups are protected (and strengthened where applicable) – while avoiding dangerous hate speech at the same time. It is therefore crucial to pay attention to the specific ethnic dynamics in society before addressing (and legislating) hate speech as well as the underlying context; otherwise political debate – particularly when it is ethnically structured – can be stifled. Yared Mengistu, for example argues that real equality amongst (predominantly ethnic) groups in deeply divided societies rests upon a censorship of hate speech that specifically targets historically oppressed minority groups in order for ruling groups to entrench their power and status.20

Furthermore, it is important to note that the scope of hate speech can change over time. In deeply divided societies, identity politics dominates and determines access to resources. As a result, parties may tend to organise along ethnic-linguistic cleavages, but later may shift to a different style of political organisation and mobilisation, which has consequences for how we are to view equality, and therefore analyse and map and legislate against hate speech as a result.

The Intention of Hate Speech

Hate speech is able to influence the behaviour of in-group members through socialisation to adopt and understand a particular ideology; to recruit members through the construction of a common enemy which is constructed as evil and/or a cultural or economic threat to the in-group. Hate speech can therefore serve as an effective tool to intimidate minorities; promote violence and intolerance; and recruit new members – and moreover, allow these messages to form part of the permanent visible fabric of society. Hate speech can be transmitted through a number of media, and can attack any number of groups.

Many definitions of hate speech include a clause which stipulates that the hate speech is a deliberate, or intended, discursive attack on a particular individual based on their membership of a particular group. However, arguably the anonymity afforded to internet users changes the space and context of hate speech, and therefore it is difficult to decipher the subjective intent of the perpetrator of the hate speech offence. Indeed, the “safe space” of various social network platforms can lead people to say things they probably would never say in the public space, and certainly would not carry out the threats they make. This disconnect has important implications.

For instance, the UK think tank Demos recently conducting some research into the use of racial and ethnic slurs on Twitter in the UK, and the findings may have sizeable implications in the way in which we see hate speech.21 The think tank found that there
are approximately 10,000 users per day of racist and ethnic slur terms in English, which equates to about one in every 15,000 tweets. However, they also found that slurs are used in a very wide variety of ways, both offensive and non-offensive, and were coded as: “negative stereotypes; casual uses of slurs; targeted abuse; appropriated; non-derogatory; offline action/ideologically driven”. Interestingly, the majority of slurs were deemed non-offensive and non abusive, but rather used to express in-group solidarity or non-derogatory descriptions. There were very few cases that presented an imminent threat of violence, or where individuals directly or indirectly invited offline violent action.

Question of balance

One of the most complex challenges in adopting any definition of hate speech is to ensure that it is broad enough to capture emerging expressions, targets, and forms of communication which offend and promote violence against a wide number of social and demographic groups. However, it is also important, particularly in the context of debates of hate speech versus freedom of expression that it also takes into adequate account the real rather than the perceived or politically charged, risks carried by those expressions and the channels used to disseminate them. Therefore, in order to effectively map and analyse the phenomenon of hate speech, it is important to consider the impact and implications of particular instances of offensive discourse. On a more pragmatic level, this has an evident impact on the tools that can be designed and used to map and understand hate speech, as well as in shaping responses to it.
SECTION 2: A METHODOLOGICAL TOOLKIT FOR ANALYSING HATE SPEECH
There are a number of methods that can be applied to analyse both the nature and transmission of hate speech. Crucially, these methodologies are shaped by the definition of hate speech used. However, as highlighted in the previous section, given the contestability of the term, there is a tension on the one hand between techniques that allow for a greater understanding of the context in which hate speech occurs, but which require time-intensive and in-depth engagement with each utterance/form of expression; and on the other hand, with approaches which capture larger amounts of data, but also risk including statements that cannot be considered expressions of hate speech.

• **Discourse analysis:** places the text in its wider political and social context in order to understand the currents of thought which illustrate and rationalise why it is to be considered hateful;

• **Content analysis:** analyses the text deemed offensive in order to pick out the key semantic components and targets of the speech act. This can then be coded and quantitative techniques applied to draw wider patterns and trends;

• **Automated techniques:** a relatively novel method of tracking hate speech that can be usefully employed to mine high volumes of text from different sources to search for key words and phrases which are likely indicators of hate speech in an efficient manner.
These approaches should not be seen to be mutually exclusive – indeed, researchers should attempt to use them in combination when mapping hate speech, particularly online, in order to build a more insightful and realistic picture into the problem, and ultimately drive responses and policies in an appropriate manner.

Discourse analysis

The practice of labelling speech as ‘hate speech’ can be contentious and often dangerous – it therefore requires a sound understanding of the context, tone, target and use of coded language, which may form part of, hate speech. The method of critical discourse analysis (CDA) allows for an understanding of the underlying political context and dominant ideologies that prevail in any given society.22

Discourse is a particular way of talking about and understanding the world and therefore when it comes to understanding something as complex and contentious as hate speech, a methodology which takes into account the use and implications of specific language is crucial. The theory and practice of CDA allows us to investigate discursive practices more closely. Language is loaded and powerful and contributes to the construction, perpetuation or dissemination of power hierarchies since it is reflective of a particular view of individual identities and social relationships; this is particularly applicable in the case of hate speech where specific linguistic canons are constructed and manipulated by actors to achieve specific political outcomes. At a more practical level, this translates into a deep exploration of the ideological underpinnings of certain tropes and text and how these are framed. Analysis of CDA also involves considerations of intertextuality – i.e. the relationships and links to be drawn between various pieces of texts, and how, more specifically, the incidence of hate speech is not isolated, but instead is situated in a certain political and social nexus.

Using tools such as discourse analysis, which relies heavily upon the examination of the wider social, and political milieu in which speech occurs may enable researchers to adopt broader definitions of hate speech. Discourse analysis allows for greater nuance in the description, explanation, and exploration of offensive statements, and enables readers and researchers to closely follow the process through which certain statements have been interpreted. When adopting research methods that lead to a numeric and quantitative elaboration of speech it may be advisable to rely on narrower and less ambiguous definitions of hate speech to avoid risks of capturing instances of offensive speech that are not actually hate speech. On the contrary, discourse analysis can be paired with broader conceptualisations of hate speech, as it allows capturing some of the subtleties, complexities – but also the contestability – within certain statements to provide a more holistic understanding into the incidence, use and intention of a speech act.
CASE STUDY: MMPZ ZIMBABWE

The Media Monitoring Project Zimbabwe (MMPZ), originally launched 1999 to promote freedom of expression and independent media, publishes monthly reports covering the incidences of hate speech in the country, aiming to offer a snapshot analysis of the media’s use of hate speech and its use by public figures. It maps the incidence of hate speech in the country across a wide range of media outlets in order to gain a useful cross section of opinion to comprehensively evaluate the news and media landscape. This initially began with traditional media, including print, radio and television, but has recently included online media to its diverse range of sources.

Previous research conducted by the MMPZ in Zimbabwe highlighted that hate speech is endemic in the mass media and in particular the arm that is government controlled. Particularly during times of elections, Zimbabwe’s media landscape sees an increase in the use of hate speech.

The MMPZ identified four key groups of perpetrators of hate speech: i) politicians; ii) state media columnists; iii) editors and media practitioners; iv) Pro-ZANU PF elements.

A number of different types of hate speech were used across various media platforms. Discourse analysis was able to contribute to the identification of hate speech, but also an assessment into the danger and implications of defamatory, prejudicial and insulting speech. For instance, the use of divisive language was underpinned by an intention to demonize those considered holding opposing views and presenting these as ‘enemies’ of the Zimbabwean population. This was effective given the salient political narratives that have existed over the course of Zimbabwean history and its liberation struggles – particularly in respect of anti-colonial discourse, the role of freedom fighters, land issues, and indigenization policies. Those holding differing opinions to the dominant ZANU PF narrative have long been portrayed as ‘sell outs’, agents of the West and ‘enemies’ of the Zimbabwean population. Therefore, when mapping the incidence of hate speech in the media and across political platforms, perpetrators were able to harness and politicise this narrative to incite violence and hatred, characterised by threatening language. For instance, statements in newspapers referred to “ZANU PF versus colonisers”, “foreign sponsored” and “puppet” parties. Other instances of hate speech also associated targeted groups of individuals with animals, dehumanising them in the process through this linguistic device.
Content analysis

Content analysis is a broad term and can refer to the analysis of various content features including text, images, and audio-visual material. Similarly to the case of discourse analysis, the ability to analyse various forms of expression is particularly important when considering the problem of hate speech, as while the term may immediately invite us to consider it as solely a speech or textual occurrence, the ways in which hate speech may be communicated are often difficult to divorce from its surrounding aesthetic “packaging”. This may be in the form of colour, picture, and moving images all found in videos, posters and even graffiti.

Content analysis provides the basis for a systematic and more quantitative based analysis of communication content through the way that it is able to identify, “code” and then amalgamate data from a number of sources in order to provide an insight into patterns and trends which can link various seemingly disparate occurrences of hate speech. For example, content analysis allows individuals to identify and count the occurrences of specific characteristic or dimensions of text which can in turn be used to identify and understand any recurrent themes.

The case study below offers a different framework to that of the MMPZ; the UMATI Project was launched in Kenya as a result of the role of hate speech in fuelling the 2007/8 post-electoral violence, and focused upon the incidence of online hate speech, especially on social networks. The methodology used for content analysis is based on the framework developed by Susan Benesch, and therefore adopts the narrower conception of “dangerous speech” as its operational principle.24
CASE STUDY: UMATI PROJECT, KENYA

Hate speech was identified as one of the key channels through which the 2007-2008 post electoral violence was fueled, in which over 1,200 people lost their lives and over 600,000 were displaced. One such example of a hate speech act from the 2007 presidential election campaign were the coded comments made by radio presenter Joshua Arap Sang on his morning radio show which broadcasts in the vernacular Kalenjin language to communicate to his listeners where and when to commit attacks on rival political party supporters.

It is widely believed that the 2007 electoral violence was facilitated by online technologies to spread inflammatory speech and threats – and in advance of the 2013 election, iHub Research and Ushahidi searched for and analysed Kenyan inflammatory discourse. Beginning in September 2013, and continuing until April 2013 or later, the Umati project monitored Kenyan online discourse in order to estimate both the occurrence and virulence of hate and dangerous speech. The project team consists of groups of human monitors, who looked for sources of hate speech online in the most prevalent languages used in Kenya: the vernacular languages of the four largest ethnic groups in Kenya (Kikuyu, Luhya, Kalenjin and Luo), Kenya’s national language Swahili, and the unofficial slang language, Sheng, which is used widely in urban centres, and Somali, spoken by the largest immigrant community in Kenya.

The technique employed to conduct this analysis requires monitors to first scan online platforms (such as twitter and Facebook) for occurrences of hate speech. They are then saved in an online database and coded accordingly. This simultaneously allows for a quantitative approach (to map how many incidences are taking place and where), but also a qualitative approach through analysing discourse. Once this was carried out, texts were translated into English and sorted into three categories: offensive speech, moderately dangerous speech, and extremely dangerous speech. In order to do this, monitors evaluated two questions which were based on the influence the speaker had in the online community, the content of the statement, and the social and historical context that the speech occurred in:

1. On a scale of 1 to 3 with 1 being little influence and 3 being a lot of influence, how much influence does the speaker have over the audience?

2. On a scale of 1 to 3 with 1 being barely inflammatory and 3 being extremely inflammatory, how inflammatory is the content of the text?
The findings of the UMATI project are varied and interesting; more specifically, they invite a questioning of the mediums in which hate speech is investigated; the importance of an adequate definition and conceptualisation of the term; as well as the role of context and intention. This is helpful in providing an informative framework which is able to inform future application in the country, as well as in other geographical units. Some key conclusions are highlighted below:

• UMATI found that the occurrence of hate speech online did not directly translate to the incidence of violence on the ground; instead other factors played a more significant part in accounting for the incidence of violent or indeed peaceful outcomes. Given this, UMATI considers online hate speech to be a window into the conversations Kenyans engage in the offline space, and therefore offer a mechanism to understand recurring issues that need to be addressed;

• Hate speech was much more prevalent on facebook than twitter – this is thought to be the case as it is quite common on twitter that tweets that are deemed unacceptable by the status quo are openly shunned, and the author of tweets subjected to ridicule;

• Definition of hate speech is important as survey data highlighted that there was a polarity between public perceptions of what constitutes hate speech, and what the UMATI project defined it as. For instance, the Kenyan public considered personal insults, propaganda, and negative commentary about politicians as hate speech. At the same time, the public held more broad conceptualisation of the targets of hate speech than the Constitution, which focuses on discrimination on tribal lines;

While stereotypical insults across tribes would fall into Benesch’s definition of dangerous speech, the intent of such insults (as well as the context of them) were usually largely perceived as harmless banter.
SMS messages are often a cheap and effective means of communicating not only with single individuals, but also in sending mass text messages to groups particularly in regions where access to the Internet is limited and expensive. Over the past decades, mobile phone use has expanded exponentially, especially in sub-Saharan Africa, and with it a new medium for the communication of hate speech. The results of the post-election enquiries conceded that hate speech and incitement to violence were also disseminated via SMS messages on mobile phones. The following are existing examples that are reportedly characteristic of typical SMS messages circulated during the 2007/8 elections:

“Fellow Kenyans, the Kikuyus [Kenyan tribe] have stolen our children’s future. Hope of removing them through the ballot has been stolen. We must deal with them the way they understand violence. We must dominate them.”

“We say no more innocent Kikuyu blood will be shed. We will slaughter them right here in the capital city. For justice, compile a list of Luos and Kalus[ph] [ethnic communities] you know at work or in your estates, or elsewhere in Nairobi, plus where how their children go to school. We will give you numbers to text this information.”

However, many communications companies, Safaricom included, maintain that they have neither the legal capacity nor the technical capacity to even store SMS messages for retrieval and analysis. Companies may also be precluded from collecting evidence of hate speech as it may compromise user privacy.

The use of mobile phones in Kenya is still increasing, and in the period from 2007-2013, mobile subscriptions in Kenya had increased from 12 million to 29.8 million. Indeed, Safaricom predicted that 2 billion SMS texts would be sent in the run up to the election, double that of 2007 – posing a threat ahead of the 2013 election. Beforehand, Kenya’s largest mobile phone operator, Safaricom, decided to develop its own practices and set up preventative messages against the dissemination of hate speech by vetting bulk SMS content that political parties, politicians and aspirants wished to send in ahead of the elections. Under the new guidelines set by Safaricom, anyone wishing to send bulk SMS messages with political content would have to first submit an application to Safaricom ready for them to vet the content of these messages to verify that they did not contain instances of hate speech. In addition to this, Safaricom applied and received the support of the government and the Communications Commission of Kenya (CCK) to develop a formal oversight framework, which applied to all mobile network operators in the country,
Guidelines for the *Prevention of Transmission of Undesirable Bulk Content/ Messages via Electronic Communications Networks* in October 2012.

The guidelines addressed a number of issues thought to have promoted the 2007 post election violence, including:

**Identifying the sender:** in the form required for submission in advance of sending bulk SMS messages, and to combat texts received from unknown numbers, applicants were required to provide the following: the verbatim content of the political message; a signed authorisation letter from the political party representative or individual sending the message; and certified copies of political party representative or individual sending the message.

**Language:** Since there are 42 languages in Kenya, and many are specific to particular ethnic groups, as well as frequent slang speech Sheng among urban youth, heavily coded messages can be used to incite hatred and deliberately exclude other communities. Incitement uttered in local languages also meant that the risk of detection was lowered. In order to combat this, Safaricom insisted in its new guidelines that content for political bulk SMS messages was sent in Kiswahili and English only.

**Content:** Safaricom would ‘vet’ the content of the proposed message in accordance with the following official guidelines: i) political messages shall not contain offensive, abusive, obscene or profane language; ii) political messages shall not contain inciting or discriminatory language that may or is intended to expose an individual or group of individuals to hatred, hostility or ridicule on the basis of ethnicity, tribe, race, colour, religion, gender or otherwise; iii) political messages shall focus on Party manifestos and shall not dwell on unnecessary attacks on individual persons, their families, their tribe or their associations. All messages would have to be personally approved and signed off by Senior Management.

**Unsolicited messages:** During the 2007/8 post election violence, some mobile phone users reported receiving up to 50 ‘unsolicited’ SMS messages per day. The Safaricom guidelines state that sending messages of a political nature can only be sent if the user has opted-in, or subscribed. In the case of reports of unsolicited messages, Safaricom would reserve the right to immediately terminate the license agreement.
Automated techniques

With the increasing proliferation of news and information sources available on the Internet, especially social media, there are a number of spaces available to the public to not only consume information, but also participate in shaping it. For example, forum discussions, blogs, Twitter, Facebook, YouTube and other social media sites all serve as a rich repository for analysis. However, social media datasets are often far too large to be manually analysed, and therefore benefit from automated techniques, which can present a number of opportunities for more complex and sophisticated data analysis. Sentiment analysis or opinion mining, for example, studies popular opinion on a wide variety of topics from products and services, to organisations and individuals, as well as specific events and topics. Moreover, specialist software packages allow for the processing of social media data at great scale and speed. Opinion mining is allows researcher to design their own bespoke filters; organise and sort data according to various categories iteratively; is able to analyse at the document and sentence level; and can also gauge positive and negative sentiment and comparative opinions.

CASE STUDY: DEMOS PROJECT - RACIST SLURS ON TWITTER

The UK think-tank, Demos, recently conducted a project, which sought to investigate the incidence and nature of racist slurs on the social media platform, Twitter. In order to do so, they used the publically available live Twitter feed as a data set, through its ‘stream’ application programming interface, and used this to search for tweets containing one or more racial slurs. Slurs were crowd sourced from Wikipedia, with the final list being edited before running the programme to compile a dataset of all offensive tweets that met the criteria. After this, the tweets were filtered to remove instances where the slur terms were part of an account name, and finally passed through an English language filter to exclude non-English tweets.
Towards a Framework for Further Analysis

The large palette of methods used to map and analyse hate speech reflect the diverse number of sources and types of hate speech that permeate both online and offline spaces. However, as evidenced above, there are a number of advantages and disadvantages associated with each approach. More specifically, there is an important trade off to be made between adopting those techniques which allow for a more nuanced analysis of content, and those approaches which are less time and labour intensive, but that are able to process high volumes of data. Innovative approaches, such as those employed by UMATI, use a combination of the methodologies of content, discourse and automated technologies and have drawn insightful conclusions which are able to shape understanding of both the nature and dissemination of hate speech. They are also able to usefully demonstrate to what effect the trade off between consideration of context and data quality has been negotiated.

A critical evaluation of the methodologies highlights the importance of designing a framework which is able to reflect commonly held definitions of hate speech amongst the groups of interest, as well as government legislation. The case studies also highlight the importance of context in evaluating the danger and intention of hate speech, which is crucial in the policy-making setting. For example, the MMPZ highlighted the relevance of historical narratives in identifying the occurrence and danger of hate speech; the Demos project used automated techniques to highlight that certain statements which could initially be characterised as hate speech were not widely perceived to be hateful or incendiary; while all the projects captured the incidence of hate speech across a diverse range of media in an appropriate and informative way. Critical evaluation of these and other approaches are also able to provide a framework and guidance to design further studies.
SECTION 3: HATE SPEECH IN ETHIOPIA
The penetration of the Internet in Ethiopia is particularly low (only 1.5% of Ethiopian citizens had access to the Internet in 2012), but the experience of the elections in 2005 has demonstrated the ability of messages originating from the Internet to be relayed through other media and reach wider audiences, especially in urban areas. The salience of the online sphere as a space for contestation has been recognized by the government of Ethiopia through the expansion of anti-terrorism laws to encompass the online space. The law prohibits “the use of any telecommunications network or apparatus to disseminate any terrorizing message” or “obscene message”, subjecting violators to a prison sentence of up to eight years.

The debate on the type of agenda new media are promoting, however, on the real or perceived ability to mobilize individuals and towards which ends, has been largely missing. The confrontation has tended to be polarized between calls to respect freedom of expression by human rights organizations, private individuals and international pressure groups – which are well meaning and legitimate but have had little effects in the Ethiopian context – and assertions by the government of Ethiopia of its right to defend national sovereignty and ensure the stability of the country. Very few attempts have been made to understand the issues that are actually discussed online and offline and to assess the risks conversations around them, as well as calls to action, may or may not carry.

Emerging from the collaboration with the School of Journalism and Communication at Addis Ababa University and from a workshop that took place in Addis Ababa in February 2014, this last section of the working paper reflects on the opportunities and challenges of analysing speech online in Ethiopia. The workshop gathered together representatives of the Ethiopian Government, of the major opposition parties, civil servants, members of international organizations and of civil society groups, and academics, as well as journalists and bloggers at different ends of the political spectrum in Ethiopia.

The uneasy relationship between the media and the state

Since it took power in 1991, the Ethiopian People’s Revolutionary Democratic Front’s (EPRDF) relationship with the media has been contentious and adversarial. In large part, this is the result of the government’s attempt to balance public pressure for freedom of expression and debate with its own desire to use the media to forward its own political agenda. After the contested elections of 2005, when the government stepped up repressive measures against independent journalists and began implementing one of the most pervasive systems for filtering online political content in sub-Saharan Africa, its handling of the media has become the target of widespread criticism from NGOs and human rights groups. Their demands for the government to change course, reform the media and guarantee greater freedom of expression, however, have largely been ignored.
Recent events have to be located in the longer history of the relationship the government developed with the media and with what can be considered the EPRDF’s “original sin” in the media sphere: the fact that when the new leaders came to power they opened the space for debate but refused to engage with the very debates they had allowed to bloom. In the early 1990s, the EPRDF had to show it was different from its oppressive predecessors; in the ostensibly unipolar world that emerged after the fall of the Soviet Union, the pressure to respect certain rights and freedoms was significant and freer media represented an opportunity to boost the new government’s international legitimacy. At the domestic level, the liberalization of the press also helped signal to a population traumatized by decades of war that a new breed of rulers was now in power. The Transitional Federal Government, which was established to write a new constitution and build the foundations for a new Ethiopian state, soon created the conditions for the first private papers to start publishing and later spelled out their rights in the relatively progressive press law passed in 1992. However, these measures were to be undermined by the EPRDF’s lack of commitment to the freedoms it had allowed, and its failure to understand what it really meant to allow a plurality of voices to compete in a post-war setting. Although the criticism took on an increasingly adversarial tone, the EPRDF leadership initially stuck to its policy, ignoring dissenting voices and labelling them as “anti-peace” and “anti-constitution”. This stemmed from a belief that those writing for the private press were not part of the EPRDF’s constituency in any case, so there was little need to expend political capital either repressing or engaging them. Over time, however, the trading of accusations and the inability of opposing factions to command each other’s attention progressively poisoned the debate in ways that would have repercussions beyond the press.

When the Internet started to be employed as a space to discuss Ethiopian politics, the debates were rapidly captured by the polarized tones that had characterized traditional press outlets. Platforms such as the Ethiopian Review, Nazret, and Ethiomedia, all launched by Ethiopians in the diaspora, hosted articles that could have equally appeared in the newspapers printed in Addis Ababa. Indeed, from the very beginning it became common to find references and connections between online and printed articles. The new media, rather than being seized by a new generation of leaders and advocates as an opportunity for to test innovative ideas, were largely captured by “old politics”. Instead of debating new issues, as occurred in nearby Kenya, authors returned to old grievances and were unable to galvanize or mobilize passions and political energy in the same way as more extreme pieces that polarized the debate. The Reporter, a newspaper started by Amare Aregawi, a former member of the Tigrayan People’s Liberation Front (TPLF), the precursor of the EPRDF, and editor in chief of TPLF’s clandestine radio during the guerrilla struggle against the Derg regime was started to provide a form of constructive criticism to the EPRDF. Papers such as Meznania were representative a younger generation trying to move the debate to the centre. Similarly moderate online spaces were those of created by bloggers such as Dagmawi or Enset.
When international organizations and NGOs began to hail the potential of new media such as the Internet for development in the second half of the 1990s, the government of Ethiopia remained largely indifferent to such ideas, even if it did not challenge them openly. In the absence of a clear understanding of how technology could be used to serve the state’s interests, and fearful that the new media could be turned against them, the government simply prevented other actors from entering the market. Telecommunications continued to be a state monopoly, no private Internet Service Provider (ISP) was licensed and the cost for connectivity was kept out of the reach of the vast majority of Ethiopians. As would become clearer during the elections in 2005, however, this form of resistance towards the new media, was not sufficient to prevent it from playing an important role in mobilizing dissent against the government. The policy towards maintaining monopoly of the Internet and mobile phones did keep their penetration very low, at 0.22 and 0.55 percent of the population respectively in 2005. However, when analysed as part of the wider communication network they became part of in 2005, it is apparent how these media had a potential that went well beyond what these numbers alone may suggest. By 2005 the Ethiopian blogosphere was blossoming. In the years preceding the elections many bloggers had not only joined already popular platforms such as Nazret, Ethiomedia, and the Ethiopian Review, but had started creating their own spaces. Bloggers such as Enset were influential commentators from the diaspora, while others like Ethio-Zagol were contributing to the online debate from Addis Ababa. As mentioned earlier, in a move that dramatically multiplied the possibilities of these voices to be heard in Ethiopia despite the limited Internet connectivity, many newspapers reproduced opinion pieces and news originally published online. The quest for information also led people to download and print news, commentaries, and political manifestos, turning them into leaflets to be distributed to those that could not have access to the Internet. Mobile phones, and especially SMS, were used to mobilize people in real time and to disseminate calls for action which had first emerged on other platforms, while the word of mouth continued to play a paramount role in further disseminating information, especially in urban settings.

After Election Day, when the EPRDF realized it had suffered greater losses than it was ready to accept and people started protesting over the delay in issuing the results, some of the channels used to mobilize protestors were shut down. In the aftermath of the first wave of demonstrations initiated by university and secondary school students on the 6th June, the SMS service was discontinued, and only restored some two years later. However, it took some time for the government to realize how opposition groups were using a variety of media, as part of a more complex network where information could be relayed across multiple channels. After shutting down the SMS messaging service, the Ethiopian government progressively continued to close down other communication channels so as to reduce their capacity to be employed, singularly or in combination, to serve the protest and disseminate alternative information and narratives. In early November 2005, some of the most vocal Ethiopian journalists who challenged the results of the election and called for more democracy were arrested and their papers were forced to close. In May 2006, one year after the contested election, access to online spaces such as Nazret and Ethiomedia, and to a number of individual blogs, started to be blocked.
Online spaces, ethnic divisions, and the use of hate speech for political ends

The 2005 elections became the fundamental juncture when the tension between politics, ethnicity, and the media, including new media, became evident. Opposing factions accused one another of resorting to ethnically charged language and read contemporary political agendas as an expression of the aspiration of competing ethnic groups to dominate over the other. Language of hate speech was also evoked as a part of this electoral rhetoric, focusing on recent examples in Rwanda. During the elections one of the most contested political issues were drawn around whether to define the Ethiopian state according to ethnic affiliations. The major opposition party, the Coalition for Unity and Democracy (CUD), advocated a return to a more unitary model of the Ethiopian state whereas the ruling party criticised CUD for trying to re-instate old Amhara dominance in Ethiopia. According to an International Crisis Group report, the rhetoric of these elections revolved around these polarising viewpoints:

The election rhetoric linked CUD agenda to that of Rwanda’s Hutu extremists (the Interahamwe) and described southern peasantry in imperial times. Nationalist politicians responded with the charge that the EPRDF was breaking the Ethiopian nation-state into ethnic groups as part of a Machiavellian divide-and-rule strategy.

Other commentators have noted that some of the political violence did replay older animosities between the Tigray and Amhara ethnicities in Ethiopia. Yared Mengistu, for instance, writes that:

[There] is an unsettling resemblance between the hate propaganda used during the Rwandan genocide and the hate campaign surrounding the May 2005 elections in Ethiopia. Fortunately, Ethiopians did not experience killings of genocidal proportions, although the election air was charged with hate, recrimination, and bloodshed. [...] This is because the top echelons of the country’s ruling party, who are predominantly of Tigrayan extract, believed that the Amharas were out to regain their long-held dominance of the political landscape. The predominantly Amhara major opposition coalition, the Coalition for Unity and Democracy (CUD), contributed to the creation of Amhara-Tigray battle imagery.

These animosities were indeed reflected in the social media. As the excerpts below from Ethiomedia and the Ethiopian Review indicate, the attacks waged against the Ethiopian government were not simply critical of its governance, but were directed at some of the pillars on which the national agenda was founded, indicating how the principle of ethnic federalism on which the nation was based and that informed the drafting of the constitution was actually meant to fuel violence among ethnic groups.
The Meles propaganda machinery and his state apparatus are tirelessly promoting all sorts of machinations aimed at turning one Ethiopian ethnic group against another. In a police state where mismanagement, nepotism and corruption are rampant, and where the people are completely disenchanted and disgusted with the authoritarian rule of a tyrant, the illusion of stability and security would remain whimsical.35

Ethnic-based parcelling of the map of Ethiopia with evil design to make it convenient for cessation. This is a reckless and dangerous experiment, which should be utterly condemned not only by Ethiopians but also by all Africans suffering from tribal strife as a cancer to peace, stability and prosperity. Creation of a feudal land tenure system in which TPLF is the landlord and the peasants live in slavery in serfdom.36

However, the heated and polemical language employed in these and similar posts were not characteristic of oppositional voices alone, but were a trait common to most of the players participating in the polarized political debate in Ethiopia. A few days before the elections, for example, the Prime Minister himself declared in a televised interview:

*I call on the people of Ethiopia to punish opposition parties who are promoting an ideology of hatred and divisiveness by denying them their votes at election on May 15. Their policies are geared toward creating hatred and rifts between ethnic groups similar to the policies of the Interahamwe when Hutu militia massacred Tutsis in Rwanda. It is a dangerous policy that leads the nation to violence and bloodshed.*37

In the months following the elections and the unrest their uncertain result triggered, Meles’ argument was often turned on its head by bloggers who used similar or greater vehemence to accuse him, and not the oppositional forces, to be the main cause of ethnic divisions.

*It is now clear that Meles’ campaign was centered on further deepening ethnic suspicion and mistrust in to the fiber of Ethiopian politics. Lacking results to show for the 14 years of his time at the helm, the PM chose to campaign on ethnic politics platform that shamelessly dared to bring the Rwandan experience in to the Ethiopian political discourse. Meles chose this platform, not only because he lacked a record to campaign on but he also believed that ethnic politics has taken root capable of delivering votes bounty. The lesson Meles will take from this election will be the need to accelerate his divisive ethnic project. If allowed, he will come back with vengeance. Venomous ethnic division and instigation of tension among the people will be the trademark of his rule for years to come. Instability will ensue, providing Meles with the rational for his future extrajudicial measures. He will use this to impress and cajole the international community. Beware, given the chance, he will do it.*38
These excerpts further exemplify how some of the most popular online spaces aimed at Ethiopians in Ethiopia and the diaspora were challenging the very core of the national discourse advanced by the EPRDF. They also suggested, in potentially provocative and harmful language, how ethnic federalism could slide into ethnic discrimination and ethnically motivated violence. Appreciation of the context in which polarising and provocative speech – which in other spaces could be conceptualised as “hate speech” – occurs, is therefore essential when building a framework to analyse and monitor its incidences and unmask some of the underlying motivations and intentions behind hate speech acts.

Towards a new research agenda

This working paper - and our pilot research that has informed it - stands against the backdrop of these many debates. Indeed the debates on hate speech can be seen as a kind of a nodal point around which many different and contested opinions about the freedom of speech and its limitations in contemporary societies revolve. The paper thus concludes by asking whether it is possible to approach these grievances that characterises communication online and offline in a different, creative (and possibly complementary) way? And what kind of a new research agenda could such work inform?

Perhaps nowhere are these debates as important to understand as in countries such as Ethiopia with significant levels of social and political tension. As we have shown in this paper, the debates on hate speech are often polarised with opposing sides accusing the other of promoting hate speech to serve their political aims. In the Ethiopian context, the most vocal criticism of the measures taken by the Ethiopian government to limit freedom of speech have come from international NGOs and activist groups. Such demands carry out an important role in highlighting the practices through which national legislation and international agenda are used to silence political dissent. The notorious anti-terrorism legislation in Ethiopia is a case point: because of its ambiguous wording, it has been
routinely used to arrests and sentence journalists. Yet however well-meaning and legitimate such demands are, they have been mostly ineffective when countries such as Ethiopia are concerned. At best they have resulted in a cold shrug from officials criticising foreign interference in national politics; at worst, they have resulted in accusations of journalists colluding with foreign agents to destabilise the country. The recent arrests of the Zone9 bloggers is a case in point: they were accused by the Ethiopian government of “working with foreign organizations that claim to be human rights activists and ... receiving finance to incite public violence through social media.”

Using hate speech to justify censorship is of course never justified. What, however, international criticism by organisations such as Reporters Sans Frontiers (RSF) and the Committee to Protect Journalists (CPJ), and the responses from the Ethiopian government to this criticism, clearly highlights is the complicated ways local and international discourse is now intertwined. On the one hand, debates on the global free flow of information and ICTs often relies on provocative messages and a rhetoric of radicalism that might not be sensitive to the complexity of political situation in different countries such as Ethiopia. On the other hand, the Ethiopian government has not taken any step towards reaching consensus or tried to bring back debates towards more constructive dialogue. Its tendency has been to dismiss rather than reach out to critical voices. Sovereignty has been affirmed first at the technical level, through the censoring of critical online spaces, and later at the discursive and legal level.

Given this polarised climate, the approach we propose for a future research agenda seeks to address some of these tensions. First of all, it recognises that in deeply divided societies where the interpretation of history is contested and the space for dialogue limited, different competing claims need to be all taken seriously. Such work should steer away - as far as possible - from easy attribution of political value (utopian or dystopian) onto different debates on hate speech or the relationship between ICTs and the state. Rather, it should draw its inspiration more from principles of conflict mediation whereby different voices are accepted as partners in situations where social antagonisms risks breaking out into violence. What is thus needed is a research agenda whose aim is to help mitigate such polarised environments rather than buying into the existing rhetoric. Second, such work needs to be based on empirical evidence, rather than existing political grievances, and use it as a base for broader debates. The pilot research we conducted in Ethiopia mapped out some of the instances of hate speech in the online spaces populated by Ethiopians inside and outside the country and illustrated that there indeed is a significant murmur of hate speech online through which different ethnic, social and political antagonisms play out. Even more crucial than the finding per se, this research was important as it enabled an open-ended debate about what are some of the real (and perceived) threats of hate speech to the Ethiopian society, and especially in light of its 2015 elections. The preliminary findings were in fact debated by a heterogeneous group of stakeholders, from representatives of the government of Ethiopia to leaders of the opposition, to bloggers on different ends of the political spectrum, and members of the international community. And while no agreement was reached as to what measures could be taken to counter such speech (or who bore the brunt of the blame for it), all participants agreed on the importance of new academic research for creating spaces of dialogue where existing grievances and opposing viewpoints could be addressed.
Building on this experience, two inter-related strands of work that could be advanced:

- Evidence-based academic research can provide a way to better understand what the real and perceived threats posed by hate speech are. Such research - and the engagement of a wide variety of stakeholders to discuss and refine it - can help dispelling some of the rhetoric around hate speech and its politicisation. Such a research can also provide an “early warning system” that allows stakeholders to identify and detect potentially harmful trends. Research should not be limited to textual occurrences of discourse but also develop a better understanding of different messages communicated through art, music, and performance. By consciously avoiding taking sides, such research can also provide a platform for mediation through which the antagonisms underpinning hate speech could be better identified and steps be taken to mitigate them. This research can be combined with traditional measures of conflict mediation and conflict resolution.

- This research can lead to innovative practices through which social media are used to counter hate speech. Initiatives such as the UMATI project in Kenya identified emerging trends on Twitter by which users themselves are marginalizing instances of hate speech and the individuals behind them. A better understanding of these mechanisms and of the possibilities to encourage them may offer important alternatives to harsher measures usually taken by governments, including increased monitoring and censorship of online communication.
APPENDIX: DEFINITION OF “DANGEROUS SPEECH”
DEFINITION OF “DANGEROUS SPEECH”41

1. THE SPEAKER

• Did the speaker have the authority, power or influence over the audience? Influence or authority need not derive from a de jure political post; cultural and religious figures and entertainers often have more influence over an audience than political officials. Some speakers control resources needed by an audience, or can deploy force. Any of these factors can render their speech more dangerous.

• Was the speaker charismatic or popular? A speaker may be seen as popular only by a subset of the audience, but those listeners may also be most likely to react to the speech. It should be noted that some cases present more than one speaker, such as when a radio talk-show host interviews a public figure.

2. THE AUDIENCE

• Who was the audience most likely to react to the speech at issue? In many cases, the audience may be large or somewhat indeterminate, e.g. “the public” or all listeners of a radio station. The analysis should focus on the audience that is most likely to react with violence in response to the speech.

• Was the speech directed primarily at members of the group it purported to describe, i.e. victims, or at members of the speaker’s own group, or both? If the latter is true, the speech is more likely to be dangerous. The primary audience is often indicated by the language or venue in which the speech was delivered.

• Did the audience have the means or capacity to commit violence against the group targeted in the speech? If an audience is unable to commit mass violence, incitement cannot succeed; and is not Dangerous Speech.

• Was the audience suffering economic insecurity, e.g. lacking in food, shelter, employment, especially in comparison with its recent past?

• Is the audience characterized by excessive respect for authority? This would make an audience more vulnerable to incitement.

• Was the audience fearful? Fear might be objectively reasonable or not; its impact may be equally large, and equally well exploited by a compelling speaker.
3. THE SPEECH ACT

- **Was the speech understood by the audience as a call to violence?** Inflammatory speech is often expressed in elliptical, indirect language, which can be variously interpreted. For this analysis, the only relevant meaning is the way in which the speech was understood by the audience most likely to react, at the time when it was made or disseminated.

- **Did the speech describe the victims-to-be as other than human, e.g. as vermin, pests, insects or animals?** This is a rhetorical hallmark of incitement to genocide, and to violence, since it dehumanizes the victim or victims to be.

- **Did the speech assert that the audience faced serious danger from the victim group?** Another hallmark of incitement, this technique is known as “accusation in a mirror”. Just as self-defense is an iron-clad defense to murder, collective self-defense gives a psychological justification for group violence, even if the claim of self-defense is spurious.

- **Did the speech contain phrases, words, or coded language that has taken on a special loaded meaning, in the understanding of the speaker and audience?** Such coded language is typical of dangerous speech. It bonds the speaker and audience more tightly together. Familiar examples of this are the phrase “go to work”, used as a code for killing during the Rwandan genocide, or the word “inyenzi” (Kinyarwanda for “cockroach”), used to refer to the Tutsi or even to the non-Tutsi who sympathized with Tutsi.

4. SOCIO-HISTORICAL CONTEXT

- **Were there underlying or previous conflicts between relevant groups?**

- **Were there recent outbreaks of violence following other examples of hate speech?** This would put both speaker and audience on notice that such speech can indeed lead to violence, thereby increasing the dangerousness of the speech.

- **Were other risk factors for mass violence present?** Such as weak democratic structures and rule of law, and structural inequalities and discrimination against a group or groups.
5. MEANS OF TRANSMISSION

- **Was the speech transmitted in a way that would reinforce its capacity to persuade, e.g. via a media outlet with particular influence or without competitors?** Other modes of transmission can be compelling in and of themselves, such as new media platforms that make members of an audience feel that they are part of a select and privileged group. Music can also increase the force or influence of a message.

- **Was the audience exposed to, or did it have access to, alternative views or sources of information?** Where there are no alternative sources of information, the impact of speech is much greater.

- **Was the speech frequently repeated, in similar form or content?** Repetition magnifies the impact of a message.
REFERENCES


6 http://www.coe.int/t/dghl/standardsetting/hrpolicy/Publications/Hate_Speech_EN.pdf

7 http://hudoc.echr.coe.int/sites/eng/pages/search.aspx?i=001-58274


10 http://www.demos.co.uk/publications/antisocialmedia


12 https://www.youtube.com/t/community_guidelines

13 https://www.facebook.com/communitystandards


21 http://www.demos.co.uk/publications/antisocialmedia

23 http://www.mmpz.org/


26 http://www.demos.co.uk/publications/antisocialmedia

27 http://www.freedomhouse.org/report/freedom-net/2013/ethiopia#.Uwc3mmJ tcY


37 Part of the speech is reported in the article ‘A critical look into the Ethiopian elections’ (2005), Sudan Tribune, http://www.sudantribune.com/spip.php?page=imprimable&id_article=9931


