

DISCUSSION TOPICS

The Engineering of Hate

Luca De Biase¹

Media Ecology Research Network – ReImagine-Europa

(published: 12/9/2026)

DOI

<https://doi.org/10.20368/1971-8829/1136471>

CITE AS

De Biase, L. (2026). The Engineering of Hate [Invited Paper].
Journal of e-Learning and Knowledge Society, 22(1), V-IX.
<https://doi.org/10.20368/1971-8829/1136471>

All happy families resemble one another, Tolstoy famously wrote, while every unhappy family is unhappy in its own way. Something similar might be said of hatred: love stories may resemble one another, but every story of hate is different.

Understanding the origins of individual hatred belongs to another inquiry. This essay is not about personal suffering, collective loneliness, or the historical forces that push certain people to the margins, nor it is about the breakdown of the social elevator that, for half a century after the Second World War, gave hope to the lives of a billion people in the West and now frustrates the expectations of just as many. Its focus is narrower: the consequences of the digital infrastructures through which the many forms of hatred are communicated.

The question is how, on those infrastructures, different stories of hatred are converted into terrifying synthetic phenomena and eventually converge upon a single expressive form: violence. In such environments, human diversity is channelled into stereotyped social behaviour through communication tools that are themselves highly standardised. Individual stories do not remain individual. They enter systems of connections, conventions and convictions.

These media are not neutral. They have a history. What do we know about that history as connected to the morphology of hate speech? Why does it matter? And what can be done to influence its direction?

The Structures of Hate

Hatred can develop only within structures that host relationships between people. Those structures are not the cause of hatred, but they form its operational media environment. There is probably a relationship between the configuration of that environment and the shape taken by hate speech. How might such a relationship be studied? And how might knowledge of it help us design digital platforms and media systems that are less hospitable to hatred?

Generalisation is difficult, of course. Not all forms of hatred are the same. The form considered here is hatred directed at categories of people, in which individuals become targets only because they are assumed to belong to a particular group. This is one reason such hatred is especially intolerable.

The presumed relationship between digital media structures and the morphology of hate speech raises a central question: is the digital environment especially conducive to categorisation? If so, does it create conditions that are particularly compatible with forms of hatred directed at categories of people?

¹ corresponding author - email: luca.debiase@gmail.com

The history of computing offers one obvious clue. Computer interfaces developed through menus, keywords and systems for classifying information according to standardised concepts. In the age of social media, this logic was powerfully extended to the interpretation of people themselves. The scale of digital social platforms encouraged a quantitative view of human beings. People were transformed into homogenised profiles, limited to a relatively narrow range of possible actions, processed by fundamentally reductive algorithms and encouraged to move from the status of persons to that of characters.

The masks they wore in order to function successfully within those environments gradually hardened. The language of relationships mediated by the dominant configurations of digital technology thus became compatible with an interpretation of people based less on the specificity of their biographies than on the accumulation of identities and categories by which they could be classified.

Categorisation, of course, is far older than digital technology. Racism makes that clear. But within the digital media ecosystem as it has developed so far, it is easier to see others through the categories used to classify them than to encounter them through the singularity of their lives.

All this has unfolded across different communication structures. Some are organised primarily around relationships between individuals. Others are designed to amplify the voice of one person before a multitude. Still others bring multitudes into contact with one another.

The Forms of Relational Infrastructure

For a long time, the structures that hosted human relationships could be reduced, at least schematically, to two main forms. During the first quarter of the twenty-first century, a third expanded dramatically. A fourth is now emerging.

The first two are the familiar “one-to-one” and “one-to-many” forms. The third is “many-to-many”: much has already been written about it, though its implications still deserve clarification. The fourth is “many-to-one”, and remains largely unexplored.

In everyday conversation between individuals, the most basic communication infrastructure, individual hatred is projected onto individual victims. It often remains enclosed within the home and finds its secret expression in domestic violence. The suffering of a person who hates from a position of strength is discharged upon someone in a position of weakness, becoming the victim’s unbearable suffering.

In “one-to-one” relationships, hatred is inflicted and endured in silence: amid complicity, injustice and helplessness. Without words.

In the world of broadcasting, by contrast, a leader directs the masses towards hatred of entire categories of people in order to consolidate power. The Holocaust stands as the most terrible monument to this structure of hatred. But it is not the only example humanity has produced.

The importance of this dimension extends beyond the spectacular efficiency of German and Italian persecution of Jews and of the other groups targeted by Nazism and Fascism. Colonialism, nationalism and religious war have all generated versions of the same structure: communication flowing from one to many, from the leader to the followers.

The rise of social media brought an extraordinary acceleration in access to debate and knowledge. It also enabled many people to communicate with many others. Within that environment, however, small-scale leaders of hatred multiplied, each able to address an audience enclosed within the same bubble.

“Many-to-many” communication thus became one of the emblematic spaces in which contemporary hatred and violence are expressed.

Today a further dimension is emerging: generative artificial intelligence used through the interface of chat. Each user may experience this technology as a machine with which to establish an intimate and highly personal dialogue. Yet the machine can respond only by processing knowledge produced and recorded by a vast number of people. From that collective material, it generates answers that appear personal.

The model resembles *Her*, Spike Jonze’s 2013 film, in which a man falls in love with Samantha, a highly expressive artificial intelligence, and believes that his love is reciprocated. Eventually, however, Samantha reveals that she is emotionally involved with hundreds of users at the same time.

Like ChatGPT, Samantha is not an individual conversing with another individual. She is a collective voice: the product of an algorithmic synthesis of the knowledge and language expressed and recorded online by billions of people. She is collective even when she addresses each individual personally.

The user projects an individual personality onto her. But she can inhabit many relationships at once with the same apparent naturalness precisely because she is made from knowledge produced by many.

What, then, happens when we move from love to hatred in the “many to one” infrastructure?

The alignment of mainstream models can clearly act as a restraint upon violent behaviour. But a model such as Grok, whose owner Elon Musk presents it as an expression of unrestricted and unapologetic freedom of speech, is deliberately less constrained. Over time, many other models will emerge, each embodying the assumptions of its designers and each capable of speaking from many to one while interpreting cultures of every kind, including those through which hatred

travels.

The most consequential examples may arise in cognitive warfare. States engaged in hostile operations can use generative artificial intelligence to produce vast quantities of messages designed to weaken an adversary's social cohesion, including messages of hatred.

"Many-to-one" structures can imitate all the others. They can simulate "one-to-one" relationships, amplify "one-to-many" communication and industrialise the production of content suited to "many-to-many" environments. Because they are simulations, they can reproduce almost any communicative condition.

In doing so, they may make it easier to construct the narrative environments in which hatred can spread.

Language as Technology

The importance of the structures through which hatred circulates is overstated when those structures are described in purely technological terms.

Even the algorithms that, in "many-to-many" environments, select apparently personalised and highly engaging messages in order to keep users on platforms for as long as possible have different effects in different historical periods and geographical contexts.

The historical, economic, social and political causes of hatred do not originate in the structures through which messages of hatred are distributed. But those structures can amplify their reach. They can also influence their language and perhaps even their alphabets.

Languages are not merely instruments of communicative exchange. They are saturated with cultural experience accumulated in the collective memory of peoples. That experience re-enters communication whenever an expression encounters the interpretive capacities of those who receive it.

Alphabets, in their different graphic forms, are bridges between sound and writing. More broadly, in other systems of signs, they connect images and writing. In every communication structure, alphabets and languages are part of the interface linking embodied human experience to mediated interaction.

The language that develops in digital environments changes according to the form of communication involved: whether between individuals who know one another, between a small number of leaders and vast audiences, between many influencers and numerous relational bubbles, or through automated generative chat.

In the last of these cases, alphabets become especially significant, at least at the level of their smallest units. Contemporary artificial intelligence generally operates through tokens: fragments of words, syllables, characters or other linguistic units whose possible

continuations are predicted statistically.

As technology and science advance, systems may be developed that operate through larger units of meaning, such as whole sentences or more explicit representations of context, increasing the ability of machines to interpret language in relation to what surrounds it.

Nothing prevents us, moreover, from imagining machines that operate at still more elemental levels when they must control robots and autonomous vehicles and therefore read the physical environment in which they act. At that level, they transform many kinds of signals - temperature readings, camera images, motion data, radar signals and much else - into machine-readable systems of signs.

These, too, can be understood as alphabets and languages: systems that make interaction between machines possible and that may be compared with scientifically developed models of reality.

The "many-to-one" communication structures typical of artificial intelligence interacting with individuals in specific physical environments will expand enormously with the spread of robotics.

In relation to hatred, one need only think of the drones and robots already conducting major operations in contemporary warfare. They are changing not only the practice of violence but perhaps also its meaning.

Should the same word be used when a drone shoots down another drone?

The algorithmic and automated mediation of human relationships therefore requires deeper reflection. It is not only a legal question of responsibility. It is not only an ethical question involving freedom of expression or the right to protection from violence. It is also an epistemological question.

How is the knowledge operationalised by algorithms and language models produced? How can training be designed so that respect for rights is built in from the outset, rather than added afterwards through improvised forms of alignment? And how can the likelihood of exposure to hatred be reduced in environments where the volume of messages is too great to be managed manually?

The Consequences of Categories

It is objectively difficult to define hate speech in operational terms that can support automated intervention and allow artificial intelligence itself to be entrusted with the defence of people against automated attacks.

In social media, within "many-to-many" structures, the automation of such defences has often increased the autocratic power of those who own and govern the platforms. The automated determination of what counts as hatred depends upon the categories, priorities and

designs of the owners of the communication infrastructure.

The same could happen in artificial intelligence if the underlying technologies remain proprietary and concentrated in the hands of a small number of oligopolists.

In that case, design, training and alignment impose upon users the values of owners and their organisations. Because oligopolists derive considerable economic benefits from their market dominance, they lobby intensely against regulation intended to protect citizens. So far, those efforts have achieved substantial success.

The values of corporate owners have therefore become, in practice, the rules governing relationships between citizens whenever those relationships are mediated by large platforms.

The categories through which oligopolists identify and evaluate hate speech determine how that speech is mitigated. This is both inadequate and unjust.

Meta's failure to respond adequately to the persecution of the Rohingya in Myanmar, which began in 2017 and continued for years, provides a particularly stark example. The relevant categories were absent from its systems, and the platforms did not adequately account for the local language. That kind of hate didn't exist in terms of Meta's algorithm. Meta consequently became an instrument through which terrible violence was organised.

What can be done?

Regulation, standardisation, public debate and many other democratically organised forms of intervention are and will remain necessary. But the development of digital platforms also points towards a practical lesson. Measures that reduce oligopolistic control over social platforms and artificial intelligence can also help prevent reality from being identified with the prejudices embedded in models.

Information diversity, interoperable platforms, and decentralised and widely distributed innovation are among the preconditions for a complex system capable of generating the antibodies needed to resist expressions of hatred.

Alternatives are entirely possible, and often highly attractive.

In artificial intelligence, growing numbers of users and companies are choosing open-weight technologies. These can support transparency, autonomy and the interpretability of automated systems. They can also make it easier to connect algorithmic operations with the meanings attributed to them in social contexts.

At this level, the response to hate speech becomes architectural.

The development of transparent and interpretable forms of artificial intelligence is a precondition for an open and collectively governed discussion of the expressive forms that produce violence. This requires a healthy alliance between science and innovative companies capable of offering alternatives to dominant firms.

The possibilities explored by researchers suggest that the available alternatives are far more numerous than the dominant technological structures currently allow us to see.

If the "alphabet" of large language models has so far been relatively close to the structure of fragments and syllables, new scientific approaches are seeking ways to generate language through larger units of meaning, such as sentences and contexts.

The analogy is imperfect, but it is almost as though the system were moving from a phonetic towards an ideographic form of writing. Such a development could create conditions for less schematic ways of processing information, more attentive to cultural context, differences in sensibility and nuances of meaning.

This brings us to the central issue in any discussion of hate speech in digital environments: the relationship between the architecture of digital media and the categorisation of people.

Interpreting individuals primarily as members of categories is one of the foundations of the most pervasive rhetorics of hatred. It is also a route towards some of the worst forms of suffering.

Those who hate a category deny the possibility that individuals may differ from the stereotypes attached to it. They attribute supposed collective responsibilities to individual people.

The consequence is straightforward: the freedom of expression of those who hate can effectively destroy the freedom of expression of those who are hated. The latter often withdraw from public debate simply to escape attack.

Categories and the prejudices attached to them are inherited from archives accumulated over time. They change slowly and operate covertly. Their influence is implicit in databases of every kind and explicit in many of the artificial intelligence systems developed by dominant companies.

The most effective way to reduce the impact of the prejudices embedded in oligopolistic platforms is to make those platforms less oligopolistic: by multiplying alternatives that are open and interoperable, transparent and open to criticism.

Hatred and violence, unfortunately, cannot be eliminated. But a balance among rights, and among the different qualities of civil discourse, is less likely to emerge in a context in which the circulation of ideas depends upon media monopolised by a small number of people who impose their own assumptions upon the collective creativity of humanity.

Society is not simply a collection of relational bubbles in which groups hate the categories and collective characteristics of others. Yet the dominant platforms have created an environment that reflects precisely this image of society.

Progress is not guaranteed by the absolute freedom of expression of the powerful. It is more likely to arise from the democratic evolution of coexistence in environments whose information diversity creates space for the expression of the weakest and least powerful as well.