Hatedemics

alda*
European Association for Local Democracy

HATEDEMICS is funded by the European Union (CERV-2023-CHAR-LITI-101143249)

HATEDEMICS INTERNATIONAL CONFERENCE

AI Against Hate and Misinformation

March, 13th 2025

REPORT

TABLE OF CONTENTS

Keywords: *hate speech, AI, misinformation, freedom of speech, empowerment, awareness, counterspeech, minorities, youngsters, learning, education, legislation, critical thinking.*

Hatedemics – The Project's Genesis

The idea of the Hatedemics project comes from the observation that there are a multitude of similar technologies used to fight either hate speech or misinformation, but these technologies have never been considered together to fight misinformation as a tool to spread hate. The idea of the project is to bring people from different backgrounds and with different expertise to use AI against hate and misinformation. The Hatedemics project is a project funded by the European Union, as fighting hate speech is part of the European political guidelines and priorities. The Hatedemics project is following this spirit and objectives.

The project objectives are to strengthen the preventive and reactive measures against hate speech and disinformation online using AI-based tools and to empower NGOs/CSOs, fact-checkers, public authorities, and youngsters as activists to effectively prevent and combat polarisation, the spread of racist, xenophobic and intolerant speech, as well as conspiracy theories. The focus is to address those phenomena at the crossroads between misinformation and hate speech in online dialogical settings while bringing together Fact-Checkers and NGO operators to help develop material for the platform and train students.

Keynote Speech on the Weaponisation of Speech – Ron Salaj, Impactskills.

Ron Salaj introduced the conference by explaining that in today's society, critical theory is under attack. We are witnessing a repressive appeal to free speech. In this context, hate speech is the automated symptom of defective social technical infrastructures, not an individual symptom.

"Is the freedom in free speech the same as the freedom to be protected from violence, or are these two different balances of freedom?

Under what conditions does freedom of speech become freedom to hate?"

What we can call the “speech mafia”, meaning a group of billionaires, are benefiting from the repression of free speech to serve their own economic interests: they are monopolising speech and using it as a weapon. The concept of freedom of speech is used to threaten democracy, it seems that these two ideologies are no longer compatible. Far-right parties are using free speech to force the democratisation of their ideologies. In this context, free speech is used as a constellation to dismantle democratic institutions everywhere in the world.

“Not long ago, if you wanted to seize political power in a country, you had merely to control the army and the police.

Today a country belongs to the person who controls communications. Communication has been transformed into heavy industry.”

This shift from power has never been so true. Nowadays, the power is in the hands of those who control the media and therefore communication, which has been transformed into a heavy industry. Within this framework, Hate Speech acts as an inverted epistemology of ignorance. Today’s propaganda is different: lies are structured as truth. **Hate speech is thus structured ignorance** (the Palestinian misinformation is an example, it’s not not knowing, it’s an active misinformation to not recognise the truth).

We can define **hate speech as a mechanism of dehumanisation of the other.** It is used to sort, extract, and label people. Hate speech holds a dangerous power, as when used by the wrong politicians, it means deportation, disappearance, etc. Hate Speech is the phenomenon, and AI the technological development.

AI used for hate speech and misinformation: it can label and score people, but is unable to catch a more complex relationship in society. It can be used to create a relationship of superiority and inferiority to produce hate (by generating offensive and humiliating content toward a community). AI can also scale and amplify some phenomenon, giving a representation of a biased norm.

Tools and Methods for the Effective Use of AI - Hatedemics Project Partners

The Platform, an innovative tool to tackle Hate Speech and Misinformation - Helena Bonaldi, Fondazione Bruno Kessler

The specific objectives of the Hatedemics project are developing and validating the **Hatedemics Platform**, a toolsuite that will integrate innovative AI tools and materials used for combating intolerance and discrimination online, designing and deploying **interactive training and educational paths** based on the platform as well as **raising awareness** on the importance of fake news at the basis of hate speech, and **empowering action** by engaging young people as activists.

The Hatedemics platform is co-created with the expertise of fact-checkers, NGOs operators for civil society members. It will be available in 5 languages (Italian, Spanish, English, Polish and Maltese).

First Objective: Analysing and Exploring Hate Speech in Social Media

This analysis is made by navigating the **network of the Telegram channels** to understand how hate and misinformation are entangled, understanding what are the main **topics** discussed in the chats via **word clouds** and other stats **and exploring** the available **messages** and filter them according to their features.

The data was collected on **telegram from relevant channels**: over 10 million messages, 1 000 chats, and 600 channels were analysed between January 1, 2022, and January 1, 2023. The focus was put on 7 targeted minorities: women, LGBTQIA+, Muslims, Jews, people with disabilities, black people, and migrants. This data was then integrated on the platform and used for **hate speech and checkworthiness detection**, topic modeling, target identification and constituting the infodemics risk index.

→ How does it work? Users can navigate the channels through the network. Channels are connected according to whether they recommend each other. Users can navigate this network to explore the different dynamics. It's also possible to explore the chat topics with word clouds that can be shown according to the channel or the topic. Finally, users can directly dive into the chat conversations using the chat table that provides information about the date of the message, the user who wrote it, the number of views, whether the message is hateful, check worthy, whether it can be linked to a specific topic and targeted minority (e.g. women, POC, etc.).

Second Objective: Train Students to Reply to These Phenomena Via Counterspeech

This was implemented in the platform in the shape of a chatbox exercise for students to learn how to respond to hate and misinformation by interacting with the chatbot.

"All communicative actions aimed at refuting hate speech through thoughtful and cogent reasons, and true and fact-bound arguments."

Counterspeech is the use of fact-based information to refute hate speech and counter hate. Each reply generated by the chat box is paired with a Ground (source on the article = fact checking) to make users understand how it is possible to answer hate with fact-checked replies.

Co-Creating the Platform for Training Against Hate and Misinformation - Samuel Allan, Maladita.es Foundation.

The Hatedemics project is centred on the creation of **the platform's target groups, youngsters** and stakeholders' training (teachers, educators, NGOs, CSOs, etc.). The co-creation process of the platform is organised around four elements: co-creation phase, piloting phase, educational resources, and multinational roll out. The final aim of the co-creation process is to get the platform to the target groups and stakeholders.

Why "co-create"?

The idea is not to create just another platform similar to what has already been done. By co-creating, the Hatedemics project aims to design something different by understanding the gaps in the field and boost the user's value in order to increase the final uptake and impact.

The co-creation has been made through **8 online co-creation workshops**, 4 focused on educational aspects and 4 focused on data collection, bringing together **98 experts**. In addition, data **collection workshops** have been conducted in each country, gathering **49 participants**, to present a data collection interface, gather feedback on the interface, and collect dialogical data that will later be used to train the model we will employ in the final platform.

Here are some of the **key results of the data collection workshops**: the platform is user-friendly even if there is room for improvement, the general content's quality in article selection and text generation must be improved to ensure reliability, avoid excessive numerical data, and create more natural responses. The translations are not always accurate, especially in Maltese, and there are some inaccuracies in automatic text selection that require manual correction. But overall, the platform is functional and pretty effective.

To complete the co-creation process, **educational workshops** (one in each country) have also been organised, featuring numerous external experts. The objectives of these workshops were to start exploring the potential application of the platform and extract some practical ideas for some prototype formation. The workshops lead to the conclusion that AI is effective across countries, interactive learning activities work better and so teachers' training is essential, however, the platform needs to be complemented by a wide educational process. In addition, we found out that cultural sensitivity needs to be taken into account, and that for ethical concerns, AI detection needs human oversight.

Defining Key Educational Goals to help learners address misinformation and hate speech:

- Key knowledge expected: understanding misinformation and hate speech and the relationship between them, understanding and identifying biases, stereotypes, prejudices, and emotions and learning how social media works and its role in amplifying disinformation or hate speech.

- Key skills expected: how to debunk online misinformation that fuels hate speech, creating new or alternative narratives to fight hate speech, how to debate effectively (and increase critical thinking) and developing prompt engineering skills (e.g., creating AI prompts).

Methodological Approach: Engaging Stakeholders and Target Groups in the Daily Use of the Advanced Hatedemics Platform -Lucía García del Moral, FUNDEA and Tristan Pertíñez, Centra.

In order to engage stakeholders and target groups in the daily use of the advanced platform, participatory research has been conducted with 68 participants, a wide range of stakeholders and within 5 countries. Each session explored the following key areas: operational definitions, knowledge of the legislative developments, approaches and public policies, controversies and issues, new tools and approaches, AI and educational approach, etc.

What functionalities could the tool have? The participatory research results

The platform must **be adapted to the social and linguistic context of each country**, the **messages must be contextualised** within a broader framework that includes the social, psychological, historical, and political origins of hate speech and disinformation. It would also be useful to have tools that allow analysis of the context of the person sending the message to provide a more appropriate response.

Possible **alternatives to direct confrontation** must be available: instead of always engaging in a discussion, participants suggested that in some situations, the tool could encourage users to reflect on their own beliefs and opinions. **Automating answers** to common questions or attacks could be helpful while staying in control. A feature to **identify and filter out malicious users**, such as trolls, or users whose interactions are not aimed at genuine dialogue should be included. The platform could be Integrated into social networks or applications for daily use.

Conclusions of the participatory research on the platform

There is a lack of consensus on an operational definition of key concepts (for example, the definition of hate speech varies with each country's legislation). Evaluating the intention of harm behind a message is difficult and subjective. There is a need to create "trusted environments" and professional alliances to ensure the platform's effectiveness, hybrid strategies between machines and humans must be developed.

The use of generative AI tools in educational contexts has different functional requirements depending on the type of intervention that needs to be considered. It would be efficient to add features to understand the context of hate speech to generate effective Counter- Narrative. Finally, it is important to mind the gaps in national legislation and the need for greater coordination in European regulatory development.

The variety of data from different groups collected through comparative methods enlightened a very complex context. Comparaison have been made between countries' data to understand the context better. This context makes it too difficult to define the boundaries of hate speech as it is still a developing research field. New European laws (such as the Digital Service Act) involve companies and private actors that need to be included in the project.

1st panel of discussion – the use of AI in Combating Hate Speech and Disinformation

The first panel discussion featured Hana Kojakovic, Project Manager, Media Diversity Institute, Lydia El Khouri, Programme Manager, Textgain-AI for Good, Marco Guerini, Project Coordinator, FBK and Maryna Manchenko, Project Manager, CESIE.

The discussion started with the explanation that there are **different expertise (fact checkers, NGOs)** used to fight misinformation and hate speech. Similarly, there are **different AI tools** developed to fight misinformation and others developed to fight hate speech. However, these tools that are specifically developed to address those phenomena have **not been considered together.** To counter hate speech and misinformation, the AI will function on the principles of **identifying and sanctioning,** by removing the problematic message (account suspension, remove the post, shadow ban).

Why AI? Manual intervention alone is not scalable as the sheer amount of hate generated is too much for human handling. Moreover, AI's contribution is very convenient to classify content even if simple word spotting is most of the time not enough. **The use of AI has some limitations**, such as censorship, and hindering freedom of speech, and cannot be applied to dangerous speech (especially when expressing emotions).

To counter these limitations, the idea is to use AI to counter hate speech by producing counter narrative, a direct intervention in the discussion to withstand hate messages using a non-aggressive textual response that offers feedback through fact-bound arguments. The counter narrative should be informative, should have a personality, and should have a particular style / structure.

The problems that we face now developing this tool is that the AI should automatically be up to date, which right now is not the case. Moreover, the AI should know how to distinguish hate speech from emotion, personal information, etc. Therefore, human review still plays a crucial role.

Tools to detect online hate, the example of Textgain

Textgain is a tool developed by the European Observatory of Online Hate (EOOH) to detect online hate and disinformation. The EOOH provides data on AI law enforcement in civil society among the EU member states to help us better understand the complicated context.

With this data collection, the EOOH created an algorithm, Textgain, that uses the lexicon to automatically detect the use of the words and phrases to show the scale and pattern of online trends. It then presents the results on dashboards featuring the topic trends, the percent of total posts by toxicity based on 10 platforms and 26 languages. EOOH dashboard functionality: it is possible to filter the data on baselines channels, community channels and tailored monitoring (create your own channel).

The advantages of this tool include the ease and efficiency of detecting hate speech, as well as the ability to report and track illegal hate content. The diversity of platforms provides a broad range of sources, which cannot be obtained through manual searches. Additionally, it allows for rapid awareness of online trends, influencers, and organizations.

Maryna Manchenko, project manager in CESIE then presented the monitoring exercise CESIE organised once a year to update the data. The monitoring exercises are on Facebook and X with a dual approach: both manual and AI, as there are pros and cons to both approaches.

What's the distinction between illegal hate speech (HS), legal hate speech, and toxicity but not hate speech?

The distinction between illegal hate speech, legal hate speech, and toxicity but not hate speech is largely dependent on the perspective of the individual using the term. Hate speech spans a broad spectrum of categories, with the label being subjective and determined by the user.

There are various definitions, including any form of hate speech, such as memes and GIFs; speech that is always discriminatory or pejorative; and speech that uses identity factors. The legal classification of hate speech also varies by jurisdiction. For example, in Italy, there is no law addressing gender-based discrimination, but there are legal provisions against disability discrimination. While laws exist against fascism, which makes it illegal, the application of hate speech laws depends on local legislation.

How do annotators ensure a diverse team, monitor hate speech in the media, and what is the current situation, particularly with the growing influence of far-right views?

It is crucial to have a diverse team of annotators, as this ensures a broad understanding of hate speech and its various categories. Currently, the team works with annotators who come from organizations with a deep understanding of hate speech and its nuances. The team also strives to work directly with individuals from communities affected by hate speech, as they possess the knowledge of specific words and phrases. This collaboration helps create a comprehensive database of hate speech terms.

Can something considered hate speech years ago would now not be considered hate speech because it has become mainstream?

The perception of something as hate speech can change over time, especially as social media evolves. What was considered hate speech in the past may no longer be viewed as such if it becomes mainstream. The influence of social media platforms, especially with figures like Elon Musk, has made it more challenging to have meaningful dialogue on these issues, and as a result, the boundaries of what is considered hate speech are constantly shifting.

2nd panel – Legal and ethical framework of the use of AI

Andrew Staniforth, Director of SAHER, Ron Salaj, Director of Research & Strategy, Impactskills and Stella Meyer, Senior Associate at H/Advisors Brussels discussed **the legal and ethical dimensions of the use of AI**, particularly regarding conspiracy theories and AI-based algorithms. With the rise of generative AI, the risks have grown, raising important **questions about who gets to decide what we see online**. There is a need to **draw a clear line between free speech and illegal speech**, and it's important to consider the entire system rather than focusing solely on the platform. Users must take a more active role and should not be seen merely as victims. **Educating users, along with a combination of laws and rules, is essential.**

A fundamental issue is recognizing that **AI is not just any technology; it imitates human behavior** and thus requires special literacy. AI inherently embeds politics, and while it can bridge divides, it also has the potential to create them.

As AI impacts democracy, we must consider the future of democracy itself. AI has a profound impact on people's understanding of complex issues such as freedom, discrimination, equality, and equity. With the advent of generative AI, we are moving from "*homo sapiens* to *homo syntheticus*", projecting ourselves into a future shaped by a post-democratic world, as described by Ron Salaj. AI will significantly influence the future of democracy.

What are the key ethical concerns when using AI to combat hate speech?

The concern is on where we should draw the line between what is offensive and what is not? While there are some clear cases and strict laws in place, such as anti-Semitism laws in Germany, this is not the case everywhere. The boundaries are often less tangible, as users operate in different contexts. AI can be a useful tool because of the scale at which content is produced, but there are situations that still require a human perspective. Human moderators, who often deal with the most extreme edge cases, face a significant impact on their mental health. This issue needs more attention and better support, as it is often overlooked in discussions.

Recent models have advanced significantly, thanks in part to techniques that involve both human and AI feedback. However, generative AI will, in the future, generate offensive content like never before. Privacy concerns also arise, as AI is becoming increasingly present in our lives.

This calls for a complete reevaluation of privacy as we know it. The "waking the mold" theory suggests that there needs to be political work to address these challenges. Over the long term, counter-narratives must evolve into acts of resistance, with humans playing an active role in this process. Security is another crucial issue: when building these tools, what happens if new biases are introduced into the documents or databases? This requires careful consideration to avoid unintended consequences.

What mechanism exists between the legal framework of AI and human rights principles?

Human rights laws have established frameworks, but the real challenge lies in enforcing them. **AI reflects the societal problems** we face because what goes into the system inevitably comes out, mirroring the issues present in society, including how we enforce human rights.

The European Court of Human Rights has dealt with many cases related to freedom of expression and hate speech. Speech itself is an act with the power to harm others, a fact recognized by the EU. It is our duty to reclaim and uphold all the laws that protect against such harm.

How can regulatory framework remain future proof given the pace of AI's evolution?

It is difficult to predict the future impact of AI, but one area for improvement is **fostering more global collaboration**. We need to think on a larger scale in terms of the ecosystem and stakeholders involved, ensuring a diverse group to regulate AI effectively.

Instead of focusing solely on how fast AI is advancing, we should reverse the question: why doesn't AI slow down? Regulation occurs within a democratic infrastructure, which is a slow process. We should **call for these technologies to slow down**, take a step back, and discuss what common rules and principles can be agreed upon. For example, industries like farming and automobiles are highly regulated, yet they still see innovation. Why should AI be treated differently?

What are the main challenges of what constitutes hate speech, and how come different perspectives?

Any form of speech that incites hate or violence should be addressed. It's also important to recognize the different types of hate, such as the ignorant person who believes they are speaking the truth but fails to understand the impact of their words, the classical liar, and others. These personalities exist in varying degrees, and it's important to differentiate between them. **We don't need to react to all forms of speech;** sometimes ignoring it is the best approach. Context, the individual, their position, the intent behind their words, and the potential impact of their speech all play crucial roles in determining the appropriate response.

Everyone has a **different understanding of issues**, and discussing them online can be challenging because it often becomes an "us versus them" situation. We need to work on bridging this divide, as doing so could help address the varying perceptions people have. **What one person finds offensive isn't always hate speech**. **Education, particularly in schools, is the best antidote to hate speech**, as it provides the foundation for better understanding and tolerance. Schools should be the primary place where this education is fostered.

3rd Panel Discussion – The Future of Fact Checking on Social Media

The last panel discussion featured professional fact-checkers and was based on the recent META decision to get rid of fact checking on social media and replace it by community notes. First, as an introduction to the discussion, David Fernandez, CTO at Maldita, explained that **fact-checking works and produces results** because it relies on facts. However, the environment is constantly changing, and decisions can shift from day to day. In the EU and US, we see significant influence in government decisions, and every decision made in Europe often has a response from the US. The question remains: will we win this battle?

Simone Fontana, Editorial coordinator of Facta added that **the safety of users** who rely on platforms for information is a key concern, alongside the issue of free speech. Fact-checkers are often accused of being "biased", as they can censor content. For example, when a post on Facebook claimed that drinking bleach would eliminate the effects of the vaccine, fact-checkers did not delete the message but added a disclaimer to indicate that the information was false. This approach led to accusations of bias, highlighting the **difficult balance between ensuring accurate information and respecting free speech.**

Kinga Klich, Head of Disinformation Counteraction Department in Demagog, completed the introduction by presenting that **95% of users do not click on information labeled as false,** showing that **fact-checking is more about education than banning or deleting content.** It aims to educate people about the truth, even as it navigates the balance between freedom of speech and the spread of hate speech. The future of fact-checking depends on the support from organizations and funding; without it, it could fade into obscurity. **It's essential that fact-checking remains independent** and not controlled by wealthy individuals, **as that could threaten its freedom**. Teaching critical thinking, especially to children who have access to the internet at a young age, is crucial to fostering a more informed generation.

Regarding Meta, **fact-checkers should have access to data and need proper funding to continue their work**. Without funding, fact-checking initiatives will struggle to survive. It's not just about publishing articles but also about **educating people through workshops** and other resources. If these efforts are made without proper support, they will be in danger of failing. As we are still a few steps behind AI, it's important to recognize that **AI is like a hammer: it can be used to build something valuable or cause significant harm**, as Kinga pointed out.

What place will human intervention take in fact checking but also with AI in general?

AI is developing so quickly that it's hard to predict what role it will play in just one year. The future of fact-checking will undoubtedly involve AI, but **the human role will remain crucial for critical thinking.** Humans are needed to guide and interpret the information, as we teach AI what to do, but ultimately, we need to show it how to do it. Exploring new ways of investing in or finding funding for these efforts will be necessary to ensure the continued success of fact-checking initiatives.

- Case Studies -

Case study 1. What do I know about Europe? Intergenerational workshop to promote media literacy against disinformation -Gemma Cortada, Diputaciò Barcelona

Diputació de Barcelona is a provincial government that represents the municipalities of the Barcelona province and provides them with services and economic resources. The province of Barcelona represents 300+ municipalities. The recent elections have shown that far right ideas are rising amongst the province population, especially among young people.

The objective of the workshop is to confront misinformation as well as promote and share the European democratic values to the participants, which were young people from 13 to 16 years old and seniors.

The methodology used was the following: through **intergenerational exchanges**, the young people teach seniors how to qualify the information online, after being themselves trained by a journalist specialised in fact checking. Mixed teams are made within which seniors contribute in teaching youngsters about Europe and the importance of democracy.

The projects' results are positive: fifteen workshops have been organised in fifteen municipalities, gathering **500+ participants** (including 147 seniors). Teachers gave positive feedback and the **participants' level of knowledge improved**. Following this encouraging first part of the project, the idea is now to extend the project to all municipalities of the province.

Case study 2. Forum Gegen Fakes, experience sharing from Germany - Charlotte Weber, Make.org

Finally, to conclude the conference, Charlotte Weber, from Make.org, presented the project **Forum against Hate's: together for a strong democracy**, that aims to initiate **a nationwide debate on handling disinformation** through a unique citizen participation format in Germany.

Echoing the topics and concepts mentioned throughout the Hatedemics conference, this project explores **ways to strengthen democratic dialogue with citizens**, the cornerstone of this initiative is the involvement of as many citizens as possible by including a wide range of the population in an innovative mass participation format, **finding better ways to manage disinformation and formulate concrete recommendations.**

Asked about fact-checking, she agrees with what had been said, adding that fact-checkers just add context to information without saying that they have the truth, and that the most challenging part was **to find sources to the information.** In the form of "forum against fakes", the project combines online participation with a citizen assembly in three phases, each phase being subdivided into an online phase and an assembly. The results of the project are impressive, as **more than 7,5 million citizens were reached on social networks all over Germany**.

MATERIALS

Case study: What I know about Europe? Intergenerational workshops to promote media literacy against disinformation - Gemma Cortada.

Co-creating a Platform for Training Against Hate Speech and Disinformation - Samuel Allan.

EEOH, 'Accessible AI for CSOs' - Lydia El Khouri.

Forum Gegen Fakes, experience sharing from Germany - Charlotte Weber.

Keynote Speech - Ron Salaj.

Methodological Approach: Engaging stakeholders and target groups in the daily use of the advanced Hatedemics Platform - Lucía García del Moral, Tristan Pertíñez.

The Platform, An Innovative Tool to Tackle Hate Speech and Misinformation - Helena Bonaldi.

The Use of AI Combating Hate Speech and Misinformation - Hana Kojakovic, Lydia El Khouri Programme Manager, Maryna Manchenko, Marco Guerini.

Tools and Methods for the Effective Use of AI - Marco Guerini.