HOW AI BOTS HAVE REINFORCED GENDER BIAS IN HATE SPEECH

*Daniele Battista**

*Jessica Camargo Molano***

Abstract

The aim of this article is to examine the issue of hate speech in the digital society, with a particular emphasis on the topic of gender and misogynistic hate speech. In this context, it seeks to present concrete examples of biases observed within such systems, considering emblematic cases such as Amazon's Artificial Intelligence (AI) recruitment tool and Microsoft's Tay chatbot. The objective is to highlight how such biases have the potential not only to perpetuate gender-based discrimination but also to exacerbate inequalities. In light of these considerations, the article ultimately arrives at a fundamental conclusion: the crucial need for a multifaceted approach to address the problem of misogynistic hate speech and its manifestations against women. This approach entails, above all, a steadfast commitment to gender equality and the promotion of social justice within the digital environment.

Keywords: Artificial Intelligence, gender bias, hate speech, misogyny.

Resumo

Como os *bots* de IA reforçaram o viés de género no discurso de ódio

Este artigo tem como objetivo examinar a questão do discurso de ódio na sociedade digital, com ênfase particular no tema do género e do discurso de ódio misógino. Nesse contexto, procura apresentar exemplos concretos de preconceitos observados em tais sistemas, considerando casos emblemáticos, como a ferramenta de recrutamento de Inteligência Artificial (IA) da Amazon e o *chatbot* Tay da Microsoft. O objetivo é destacar como tais preconceitos têm o potencial, não apenas de perpetuar a discriminação com base no género, mas também de agravar as desigualdades. Perante estas considerações, o artigo chega a uma conclusão fundamental: a necessidade crucial de uma abordagem multifacetada para enfrentar o problema do discurso de ódio misógino e suas manifestações contra as mulhe-

* Department of Political and Social Studies, University of Salerno, 84084 Fisciano, Italy.
Postal address: Via Giovanni Paolo II, 132, 84084 Fisciano SA, Italia.
Electronic address: dbattista@unisa.it

** International Telematic University Uninettuno, 00186 Rome, Italy.
Postal Address: Corso Vittorio Emanuele II, 39, 00186 Roma RM, Italia.
Electronic address: j.cavalagliocamar@students.uninettunouniversity.net

res. Esta abordagem envolve, acima de tudo, um compromisso firme com a igualdade de género e a promoção da justiça social no ambiente digital.

Palavras-chave: Inteligência Artificial, viés de género, discurso de ódio, misoginia.

Resumen

Cómo los *bots* de IA han reforzado el sesgo de género en el discurso de odio

Este artículo tiene como objetivo examinar el problema del discurso de odio en la sociedad digital, con un énfasis particular en el tema del género y el discurso de odio misógino. En este contexto, busca presentar ejemplos concretos de sesgos observados en tales sistemas, considerando casos emblemáticos como la herramienta de reclutamiento de Inteligencia Artificial (IA) de Amazon y el *chatbot* Tay de Microsoft. El objetivo es destacar cómo estos sesgos tienen el potencial no solo de perpetuar la discriminación de género, sino también de agravar las desigualdades. A la luz de estas consideraciones, el artículo llega a una conclusión fundamental: la necesidad crucial de un enfoque multifacético para abordar el problema del discurso de odio misógino y sus manifestaciones contra las mujeres. Este enfoque implica, ante todo, un compromiso firme con la igualdad de género y la promoción de la justicia social en el entorno digital.

Palabras clave: Inteligencia Artificial, sesgo de género, discurso de odio, misoginia.

1. Introduction

Hate speech is a form of expression that occurs in various social contexts, including political debates, artistic expressions, professional sports, and work environments. This extreme form of communication poses a significant challenge in its understanding and management, especially in the context of rapidly evolving digital technologies, particularly social media platforms. The term in question refers to speeches or messages that spread hatred, discrimination, prejudice, or violence towards an individual or group based on characteristics such as race, ethnicity, religion, gender, sexual orientation, or gender identity (Cohen-Almagor 2011). Among the pioneers of hate studies, a notable author to mention is Matsuda (1989), who developed a definition of hate studies primarily focused on discursive content. According to this definition, for a speech to fall into the hate studies category, it must present elements of racial discrimination (such as asserting racial inferiority), be persecutory, hateful, and degrading, target historically oppressed groups or members of such groups, and derive from a clear intention by the communicator to harm the target. In general terms, it is evident that such speeches can have serious consequences for the victims and contribute to the marginalization and exclusion of disadvantaged groups (Moran 1994). Furthermore, despite this field of study being more than two decades old (Duffy 2003), many questions still need to be answered. The phenomenon is inherently complex and presents significant challenges in its understanding, especially considering the apparent simplicity with which the term is used in current discourse. Despite efforts, there is cur-

rently no universal and shared definition of hate speech. This means that when addressing this concept, it is not automatic to have a clear understanding of its boundaries and distinguishing characteristics. Its characterization is a point of intellectual dispute among different worldviews, many of which are external to the Western universe and less known. Hate speeches represent a threat to social cohesion and peaceful coexistence as they promote hatred, discrimination, and the marginalization of vulnerable groups. Its manifestations can be conveyed through various means of communication, including public speeches, mass media, and increasingly through social media platforms.

This has made hate speech a particularly significant challenge in the digital age, as messages of hate can spread rapidly and reach a wide audience. Over the years, it has also taken on different meanings depending on the historical, political, and geographical context, to the point of becoming something generic and ill-defined, often used merely as a slogan. Indeed, several studies have shown that identifying hate speech on social media is not a straightforward exercise (Miranda *et al*. 2022). Individuals who spread hatred often use a series of tricks to mask their statements and make their discriminatory or violent positions more acceptable. These tricks allow them to avoid accountability and reach a larger audience, thereby perpetuating the harmful effect of hate speech. For example, haters may use irony, humour, and satire to disguise a violent narrative (Schwarzenegger & Wagner 2018).

Hate speech in the digital sphere takes various forms and uses multimedia formats to reinforce negative stereotypes through toxic language. This online environment facilitates the rapid spread of discriminatory and prejudiced messages to a wide audience. The phenomenon assumes a connection between offline and online realms, where individuals actively express their opinions and emotions in a personalized communication context. This phenomenon fits into the framework of a multidimensional reality that develops concurrently both online and offline (Boccia Artieri 2012) and highlights the online viral capacity, migrating with ease from one platform to another (López-Paredes & Di Fátima 2023).

The European Union's regulation on hate speech is aimed at preventing and countering the dissemination of discriminatory, offensive, or hate-inciting content both online and offline. EU laws and directives regarding hate speech have been adopted at both the community and national levels, and member states are required to implement appropriate measures to ensure their enforcement. One of the key instruments in this regard is Directive 2000/31/EC on electronic commerce, which provides a legal framework for electronic information services, including social media. According to this directive, online service providers cannot be held liable for content posted by users, provided that they act promptly to remove or disable access to illegal content once they become aware of it. However, the EU has adopted additional regulatory tools to address the issue of hate speech. In 2016, the Recommendation on the removal of illegal content online was approved, urging

social media platforms to take more effective measures to identify and remove hate speech content within a defined timeframe. Furthermore, in 2021, the Directive on the accessibility of websites and mobile applications of public sector bodies was adopted. This directive requires that websites and applications of public institutions be accessible to all individuals, including those with disabilities, and imposes specific standards to ensure online content accessibility. Beyond EU regulations, each member state also has its own national legislative framework to address hate speech. Therefore, specific laws and sanctions may vary from country to country within the European Union. It is important to note, however, that the regulation of hate speech must balance the need to protect freedom of expression with the need to prevent the spread of harmful and discriminatory content.

A controversial aspect emerges regarding the responsibility to regulate this phenomenon. The responsibility for regulating hate speech is a complex issue involving various entities, including governments, social media platforms, and society as a whole. Governments are responsible for creating regulatory frameworks that protect fundamental rights while defining legal standards and sanctions, as well as promoting awareness. Social media platforms must effectively address hate speech online by removing discriminatory content and implementing preventive measures, often in collaboration with governmental authorities. These platforms can employ artificial intelligence algorithms and human moderators to monitor and manage content, as well as collaborate with governmental authorities to address cases of hate speech that violate the law. However, society also plays a critical role in promoting a culture of respect, tolerance, and inclusion alongside these entities. It is essential for citizens to be aware of the importance of civil and respectful dialogue, to condemn hate speech, and actively engage in promoting the values of equality and diversity. Additionally, civil society organizations, educational institutions, and media can play a significant role in educating people about the importance of peaceful coexistence and countering hate speech through awareness and education.

In summary, the regulation of hate speech is a shared responsibility among governments, social media platforms, and society as a whole. Through synergistic collaboration, it is possible to develop a comprehensive approach to address hate speech and create an inclusive and respectful environment both online and offline. The discussion thus far has shed light on the significant role played by social media platforms, which have become increasingly central with ever more innovative tools (Battista 2023). For this reason, in order to obtain a comprehensive and in-depth overview of the subject under consideration, it is now necessary to focus attention on hate speech within the digital society. This topic is of fundamental importance as it underscores how hate speech can find expression and dissemination through digital means, often assuming relevance in relation to gender issues. In the subsequent part of this paper, we will delve into the analysis of hate speech in relation to gender within artificial intelligence and digital applications, examin-

ing how such systems can exhibit gender biases and how this can contribute to the perpetuation of gender inequalities and discrimination. Understanding these aspects is of considerable significance as it provides a comprehensive framework for the mechanisms through which hate speech manifests itself and spreads in the digital society, paving the way for further research and actions aimed at combating this phenomenon.

2. Hate speech in the digital society

New technologies and social media have revolutionized human communication and social dynamics, transitioning from vertical narration to horizontal interaction. This has given rise to the Platform Society (Van Dijck, Poell & De Waal 2018), where the internet and specialized social networks play a pivotal role in decision-making and democratic practices. Information is now created, distributed, and consumed interactively, allowing individuals to actively engage in content production and sharing. Digital platforms, such as social media, act as intermediaries, facilitating connections and the rapid exchange of ideas, opinions, and information. This transformation emphasizes the importance of dynamic and unpredictable interactions within the digital environment, shedding light on mechanisms and consequences in the digital political sphere and providing insight into decision-making processes and democratic dynamics.

The transformation described above stimulates careful and in-depth debate on issues of discussion, allowing broader engagement and inclusive participation. The complete immersion in the society of connection has significantly facilitated old and new forms of abuse (Gagliardone 2019). Moreover, it is obvious that hate speech is diversely spread on social platforms, and its dissemination occurs at an extremely high speed, which can have a significant impact on individuals' behaviour, transcending the spatial and temporal boundaries in which it originated. On the other hand, individuals can now establish virtual social connections that surpass geographical and temporal boundaries, allowing them to interact and exchange information with others instantly and without geographical restrictions, fostering a new mode of participation within cyberspace (Vesnic-Alujevic 2012). In doing so through the use of social media, those who spread hatred and aggression can find refuge in anonymity, enabling them to freely express their negative ideas without being identified. Furthermore, these virtual platforms offer them the opportunity to connect with individuals who share a similar mindset, creating a sort of community that supports and reinforces their aggressive tendencies in the name of proselytism.

According to the Anti-Defamation League's "Online Hate and Harassment" report (2020), the increasing visibility of hate speech in cyberspace represents a significant concern. The report highlights how, since 2018, there has been an

uncontrolled escalation of such speeches, and these results can be attributed to the connection between the online and offline environments, indicating that the messages disseminated on social media are intrinsically linked to the behaviours society has experienced so far in traditional media (Olmos *et al*. 2020). In this regard, a fundamental consideration arises, as this operation constantly takes place within a broad unified environment: the digital context. The synergistic interaction, which Giglietto and Selva (2014) identify as the dual-screen conception, represents an event that goes beyond mere information sharing, transforming into a complex process of data and knowledge exchange among multiple and diverse actors. This practice is characterized by the simultaneous consumption and active participation in multiple sources of information, spanning across different devices and digital platforms. Such interconnection of information and interactions represents an advanced form of engagement in the contemporary media ecosystem. Another aspect to consider in this complex and controversial phenomenon is that social media platforms particularly facilitate hate crimes among the younger audience (Valerio 2022).

Generation Z (comprising those born from 1995-2010 onwards) is the one that has had access to the internet since birth, and their first socialization with the medium revolves around the internet: Instagram, WhatsApp, Snapchat, TikTok are the daily bread of digital natives. Digital platforms are constantly working to combat the spread of undesirable content, particularly harmful and abusive comments, videos and reactions, and are compelled to dedicate significant efforts to monitor and prevent such phenomena on a daily basis (Miró-Llinares & Rodríguez-Sala 2016). However, it's challenging to stop hate speech from going viral, as even a single offensive comment or post can trigger a chain reaction of sharing and replication, intensifying the spread of hate speech and discriminatory stereotypes (Cabo & García 2017). A sort of butterfly effect finds full application in the context of hate speech on social media. According to this principle, even a small action or insignificant event can trigger a series of unforeseen and far-reaching consequences. In the context of hate speech on social media, this means that even a single offensive comment or an inflammatory post can unleash a chain of events that amplifies the spread of such harmful content.

This phenomenon underscores the importance of careful moderation and timely prevention of hate speech on social media. Despite the absence of a clear definition of hate speech, it is a growing concern, especially in online spaces that have become hostile and inhospitable, hindering free expression and public discourse. This transformation poses a significant challenge to contemporary democracies, as it threatens their functioning by diminishing democratic participation, diversity of opinions, and the creation of a healthy public sphere.

This situation is often viewed as an environmental threat, gradually eroding the social fabric by hindering well-intentioned individuals from contributing to the common good. The hostility of online environments can stifle free expression,

causing fear of retaliation or discrimination. This reduces the diversity of voices and hampers public discourse and democratic consensus. Moreover, this negative online transformation can spill over into offline interactions, exacerbating societal tensions and divisions by promoting hate speech and harmful content, ultimately intensifying polarization and intergroup tensions.

This undermines social cohesion and the sense of common belonging, which are fundamental elements for the functioning of democracies. To effectively address this challenge, however, it is increasingly necessary to adopt multidimensional approaches that involve both digital platforms and users. Platforms must take responsibility for monitoring and moderating content, implementing robust policies to counter hate speech and abusive behaviour. At the same time, users need to be aware of their role in maintaining a healthy and inclusive online environment. This entails actively engaging in countering verbal attacks and online hate through constructive responses, reporting inappropriate content, and promoting respectful and informed dialogue.

3. Artificial Intelligence: Bias and hate speech

Artificial Intelligence (AI), which has turned out to be an increasingly well-known and used device, is the subject of much study and research. On the one hand, it has the potential to change the manner of examining statistics and making choices, but on the other it fuels concerns about biases and discrimination that emerge while it is used. According to Camargo Molano and Cavalaglio Camargo Molano (2021), AI systems are independent as statistics teach, but, if the information used to educate an AI device is biased, the system may be biased.

AI bias can inadvertently amplify hate speech when the algorithms, which power AI systems, are trained on biased or unfiltered data containing hate speech or discriminatory language. It is possible to identify different types of bias that influence hate speech:

1. Data bias: Data bias occurs when the training data used to develop an AI model contains biased content or hate speech, leading the model to learn and reproduce those biases (Noble 2018). For example, in a study by Bolukbasi *et al*. (2016), it was found that word embeddings, a popular natural language processing technique, can exhibit gender and ethnic biases due to the biased nature of the training data. If a social media algorithm is trained on data that includes hate speech, it may inadvertently promote or amplify such content, as observed in some cases.

2. Algorithmic bias: Algorithms can also exhibit bias in the way in which they process and prioritize information. For instance, if an algorithm is designed to maximize users' involvement, it may prioritize controversial or extreme content,

including hate speech, because such content tends to generate more reactions and interactions from users (Tufekci 2018). An example of this is YouTube's recommendation algorithm, which has been criticized for promoting extremist content in an effort to keep users engaged on the platform.

3. Lack of context: AI systems often lack the ability to understand the context in which language is used. This situation can lead to cases where hate speech is not properly identified and filtered out, or where non-hateful content is mistakenly flagged as hate speech (Gillespie 2017). For example, a content moderation algorithm may fail to recognize the difference between a news article reporting on hate speech and a post that is promoting hate speech.

4. Feedback loops: AI systems can create feedback loops that reinforce and amplify biases. For example, if a social media algorithm is biased towards promoting hate speech, users who engage with that content may be shown more of it, creating a loop that amplifies the spread of hate speech (Milano, Taddeo & Floridi 2020). An example of this is the way in which social media platforms can create "echo chambers" where users are only exposed to content that aligns with their existing beliefs, potentially radicalizing them further.

A point wherein AI bias is especially problematic is hate speech popularity. Hate speech is a complex and multifaceted hassle, and its detection requires a nuanced understanding of language and context. However, many AI systems for hate speech detection rely upon techniques totally based on easy keywords that could generate fake positives and fake negatives. As a result, researchers are exploring extra sophisticated ways of identifying hate speech that reconstruct the broader context wherein the language is used. For instance, in an analysis published in the *Journal of Language Aggression and Conflict* (Vilar-Lluch 2023), the researcher sought to identify language processing techniques related to the language of aggression. She determined that hate speech regularly consisted of derogatory, threatening, and dehumanizing language. However, even those sophisticated methods of figuring out hate speech do not show clear evidence of bias.

In an article published in the *Proceedings of the International AAAI Conference on Web and Social Media*, Davidson *et al*. (2017) claimed that hate speech detection systems knowledgeable on facts from social media structures were able to flag more posts written with the hate language of African Americans than with the hate language of white Americans. This suggests that AI structures may mirror biases inherent in the information with which they are trained, although they are designed to be impartial.

In addition to hate speech, AI bias can also have severe implications in other fields such as healthcare and criminal justice. For example, a study published in *Science* highlighted that an AI device used to identify which patients should benefit from more healthcare assets showed bias against black patients (Obermeyer *et al*. 2019). Similarly, a research study published in the journal *Science Advances* dis-

covered that an AI machine used to study the odds of criminal recidivism showed bias against black defendants (Angwin *et al*. 2016).

These examples underscore the importance of managing bias in AI. As Camargo Molano and Cavalaglio Camargo Molano (2021, 162) state, bias in AI "can lead to faulty results, if these systems are used in social research; moreover, they stress some issues of epistemological nature."

To cope with bias in AI, it is not important to know more about the technical components of AI development, but to be aware of the social and moral implications of these structures. As a study published in the journal *Nature Machine Intelligence* notes, AI systems are not impartial devices, but rather replicate the values and biases of their creators and clients (Holstein *et al*. 2019).

To address such problems, researchers have proposed several techniques to reduce bias in AI. One approach is to use diverse and representative statistics tools to train AI systems. As Camargo Molano and Cavalaglio Camargo Molano write, there are several tools that can help to mitigate the danger of bias by ensuring that the AI device is exposed to a large number of examples and perspectives (2021). Another technique is to apply an explainable AI, which lets researchers recognize how an AI tool makes decisions and to discover any biases it may have.

In conclusion, it is essential to be aware that AI has both the potential to revolutionize studies and the capability of bringing out bias and discrimination. As researchers continually deploy increasingly innovative strategies in AI, it is going to be vital to make sure that those systems are designed to be as unbiased as possible. As Camargo Molano and Cavalaglio Camargo Molano argue (2021), the improvement of AI systems that are free from bias and discrimination is essential to make sure that those systems are used ethically and responsibly.

4. Artificial Intelligence: Misogynistic hate speech

Artificial Intelligence has proven to be an increasingly popular device in research, with the capability to revolutionize the way we examine statistics and make selections. However, the use of AI also raises concerns about bias and discrimination. One area where AI bias is particularly complex is hate speech against women. Misogynistic hate speech is a complex issue, and its detection requires a nuanced understanding of language and context. However, many AI structures used to detect hate speech rely upon techniques totally based on easy keywords that could generate false positives and false negatives.

Gender bias in AI is well documented. Waseem and Hovy (cited by Davidson *et al*. 2017, 514) observed that sexist and derogatory terms towards women are often regarded merely as offensive and not necessarily as hate speech. This implies that automated hate-speech detection systems might not be adequately tuned to capture the full spectrum of sexist expressions with the same precision as they

identify other forms of hate speech. This shows that AI structures are able to replicate the biases inherent in the statistics they train with, even if they are designed to be independent.

The problem of gender bias in AI has been developing in recent years, particularly in the context of hate speech against women. One example of gender bias in AI is the case of Tay, a chatbot created by Microsoft in 2016. Tay was designed to learn from interactions with Twitter customers and responded in a conversational way. However, within hours of its release, Tay started posting misogynistic and racist tweets, together with expressions such as "f***ing hate feminists and they need to all die and burn in hell" (Vincent 2016). This was due to the fact that Tay had been trained on a dataset of tweets that included a substantial quantity of hate speech and offensive language.

Another example of gender bias in AI is the case of the AI recruiting device, which was designed by Amazon to analyse curricula and identify the applicants with the highest certification. However, the system became biased against women, because it had been trained on a dataset of curricula submitted predominantly by men. As a result, the tool penalized the curricula that contained phrases including "women's," and even "downgraded graduates of two all-women colleges" (Dastin 2018).

Misogynistic hate speech is widespread and difficult to detect because it can take many different forms. For instance, a study published in the journal *Feminist Media Studies* discovered that misogynistic hate speech on Twitter regularly takes the form of "gaslighting," causing women to question their very own experiences and perceptions (Edwards, Philip & Gerrard 2020). It is difficult to detect this kind of hate speech by using keyword-based approaches, because it often does not incorporate specific derogatory phrases.

Reinforcing gender bias through algorithms is a cause of concern because it can help normalize anti-feminist hate speech. When users receive biased information that reflects and reinforces their pre-existing beliefs, they tend to maintain those beliefs. This is particularly troubling in the context of misogynistic hate speech because it can normalize harmful attitudes and behaviours towards women.

Numerous studies have well-documented the emotional toll inflicted upon victims of misogynistic hate speech (Henry & Powell 2018). This form of hate speech, often personalized and targeted, evokes a range of emotions, including fear, anger, and sadness. Consequently, victims often experience feelings of isolation and helplessness, with adverse effects on their mental well-being. Prolonged exposure to such hate speech can also lead to heightened levels of anxiety and depression, compounding the emotional distress experienced by victims.

Exposure to misogynistic hate speech significantly affects a woman's self-esteem and self-worth (Kearl 2010). The derogatory and belittling nature of this rhetoric can lead to negative self-perception, impacting various aspects of a woman's

life, including personal relationships and professional aspirations. The erosion of self-esteem can also make it challenging for victims to confront and speak out against such hate speech, perpetuating a cycle of silence and continued victimization (Jane 2014).

The literature highlights how misogynistic hate speech contributes to the marginalization and exclusion of women from online spaces (Jane 2016). The fear of being targeted by hate speech can deter women from participating in online discussions, sharing their opinions, or even being present on certain digital platforms. Limited online participation can have significant implications for women's involvement in public discourse and their ability to advocate for their rights and interests.

The prevalence of misogynistic hate speech on digital platforms perpetuates harmful stereotypes and discriminatory attitudes towards women, reinforcing existing gender inequalities. Hate speech contributes to a culture that devalues and disempowers women, leading to far-reaching consequences that affect women's opportunities and societal status (Henry & Powell 2018).

The literature also deals with cases where misogynistic hate speech escalates to threats of physical violence (Citron 2014). These threats can profoundly impact a woman's sense of safety and security, both online and in the real world. The fear of physical harm can also lead to self-censorship, as women may be afraid to express their opinions for fear of retaliation (Henry & Powell 2018).

The effects of misogynistic hate speech can extend into a woman's professional and personal life. Online harassment can lead to professional setbacks, such as loss of job opportunities or damage to one's reputation. Moreover, the stress and emotional toll of dealing with hate speech can strain personal relationships and negatively affect one's social life. The cumulative impact of these consequences can be devastating, affecting every aspect of a woman's life (Vogels 2021).

To overcome such problems, some researchers are experimenting with more sophisticated techniques that reconstruct the broader context in which the language is used. For example, in a study published in the *Social Science Computer Review*, some researchers reveal that they used a system learning approach to find out the language patterns that can be related to misogynistic hate speech (Kulshrestha *et al*. 2017). They found that misogynistic hate speech is frequently characterized by sex-related words, derogatory terms, and threats of violence.

However, regardless of these strategies in place, misogynistic hate speech detection remains a challenge. This is partially due to the fact that misogynistic hate speech is often embedded in cultural norms and attitudes which are difficult to determine. As a result, dealing with hate speech against women requires going beyond AI structures and technical solutions.

5. Conclusions

In conclusion, misogynistic bias in AI is an intricate and multifaceted problem that requires the knowledge of a nuanced language and context. It should be remembered that AI has both the capability of revolutionizing research and many different areas and the capability of bringing out bias and discrimination, especially in the field of hate speech against women.

Gender bias in AI is well documented in studies of AI systems used to analyse curricula where hate speech against women is often used. This situation can be attributed to the fact that AI structures are trained on datasets that reflect the biases inherent in society. Even if AI structures are designed to be independent, they are able to replicate the biases inherent in the data which they are trained on.

To overcome such problems, researchers have proposed some techniques to reduce bias in AI. One approach is to use numerous data units to train AI structures. This can help mitigate the danger of bias as it ensures that AI systems are exposed to a wide range of examples and perspectives. Another technique is to apply an explainable AI, which lets researchers recognize how an AI device makes choices and to identify any biases that can be present.

However, despite these strategies in place, detection of misogynistic hate speech remains a challenge. Misogynistic hate speech is subtle and difficult to detect because it can take many different forms. To cope with this problem, some researchers are exploring extra state-of-the-art techniques taking into consideration the broader context in which the language of aggression is used. These strategies involve a system learning approach to identify the language patterns that are related to misogynistic hate speech.

Dealing with hate speech against women requires going beyond AI structures and technical solutions. It requires a commitment to gender equality and social justice, as well as the will to apply broader cultural norms and attitudes. Using increasingly sophisticated techniques for hate speech detection and broadening the context in which the language of aggression is used will help mitigate the risk of bias and discrimination in AI structures.

To effectively tackle the issue of gender bias and misogynistic hate speech on digital platforms, it is necessary to develop and implement regulations that specifically target the problem. Such regulations should include provisions for the monitoring and auditing of AI algorithms to ensure that they do not perpetuate gender biases or amplify hate speech (Cath *et al*. 2018). Moreover, mechanisms should be put in place to hold digital platforms accountable for the content they allow to be disseminated. These mechanisms would ensure that platforms take active steps to hinder the spread of misogynistic hate speech and other forms of online harassment.

Roles of authors

DB: Conceptualization, formal analysis, research, design of methodology, writing of original draft, revision of the draft.

JMC: Conceptualization, formal analysis, research, design of methodology writing of original draft, revision of the draft.

Conflict of interests

The authors declare that they have no financial or personal interest in the research study titled "How AI bots have reinforced gender bias in hate speech" and its publication. Moreover, they declare no financial or personal relationships with any individuals or organizations that could inappropriately influence their work or the interpretation of the results of this study or of this publication.

References

Angwin, Julia, *et al*. 2016. "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Anti-Defamation League. 2020. *Online Hate and Harassment Report: The American Experience 2020*. Retrieved 25 September 2022, from https://www.adl.org/online-hate-2020

Battista, Daniele. 2023. "For better or for worse: politics marries pop culture (TikTok and the 2022 Italian elections)." *Society Register* 7(1): 117-142. DOI: https://doi.org/10.14746/sr.2023.7.1.06

Boccia Artieri, Giovanni. 2012. *Stati di connessione. Pubblici, cittadini e consumatori nella (Social) Network Society*. Milano: FrancoAngeli.

Bolukbasi, Tolga, *et al*. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." *Proceedings of the 30th International Conference on Neural Information Processing Systems,* edited by Daniel D. Lee, 4356-4364. Red Hook, NY: Curran Associates. DOI: https://doi.org/10.48550/arXiv.1607.06520

Cabo Isasi, Alex, & Ana García Juanatey. 2017. "Hate speech in social media: a state-of-the--art review." Available at https://ajuntament.barcelona.cat/bcnvsodi/wp-content/uploads/2017/01/Informe_discurso-del-odio_ENG.pdf

Camargo Molano, Jessica, & Jacopo Cavalaglio Camargo Molano. 2021. "Criticalities and Advantages of the Use of Artificial Intelligence in Research." In *Handbook of Research on Advanced Research Methodologies for a Digital Society*, edited by Gabriella Punziano & Angela Delli Paoli, 161-175. Hershey, PA: IGI Global. DOI: https://doi.org/10.4018/978-1-7998-8473-6.ch011

Cath, Corinne, *et al*. 2018. "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach." *Science and Engineering Ethics* 24(2): 505–528. DOI: https://doi.org/10.1007/s11948-017-9901-7

Citron, Danielle K. 2014. *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press.

Cohen-Almagor, Raphael. 2011. "Fighting Hate and Bigotry on the Internet." *Policy & Internet* 3(3): 1-26. DOI: https://doi.org/10.2202/1944-2866.1059

Dastin, Jeffrey. 2018. "Amazon scraps secret AI recruiting tool that showed bias against women." *Reuters*. Available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davidson, Thomas, *et al*. 2017. "Automated hate speech detection and the problem of offensive language." In *Proceedings of the International AAAI Conference on Web and Social Media* 11(1): 512-515. DOI: https://doi.org/10.1609/icwsm.v11i1.14955

Duffy, Mary Elizabeth. 2003. "Web of Hate: A Fantasy Theme Analysis of the Rhetorical Vision of Hate Groups Online." *Journal of Communication Inquiry* 27(3): 291-312. DOI: https://doi.org/10.1177/0196859903252850

Edwards, Lee, Fiona Philip, & Ysabel Gerrard. 2020. "Communicating feminist politics? The double-edged sword of using social media in a feminist organisation." *Feminist Media Studies* 20(5): 605-622. DOI: https://doi.org/10.1080/14680777.2019.1599036

Gagliardone, Iginio. 2019. "Defining online hate and its 'Public Lives': What is the place for 'extreme speech'?" *International Journal of Communication* 13: 3068-3086.

Giglietto, Fabio, & Donatella Selva. 2014. "Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets." *Journal of Communication* 64(2): 260-277. DOI: https://doi.org/10.1111/jcom.12085

Gillespie, Tarleton. 2017. "Regulation of and by platforms." In *The SAGE Handbook of Social Media*, edited by Jean Burgess, Alice Marwick, & Thomas Poell, 254-278. London: SAGE. DOI: https://doi.org/10.4135/9781473984066

Henry, Nicola, & Anastasia Powell. 2018. "Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research." *Trauma, Violence, & Abuse* 19(2): 195-208. DOI: https://doi.org/10.1177/1524838016650189

Holstein, Kenneth, *et al*. 2019. "Improving fairness in machine learning systems: What do industry practitioners need?" In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York: ACM. DOI: https://doi.org/10.48550/arXiv.1812.05239

Jane, Emma. 2014. "Back to the kitchen, cunt: Speaking the unspeakable about online misogyny." *Continuum: Journal of Media & Cultural Studies* 28(4): 558-570. DOI: https://doi.org/10.1080/10304312.2014.924479

Jane, Emma. 2016. "Online misogyny and feminist digilantism." *Continuum* 30(3): 284-297. DOI: https://doi.org/10.1080/10304312.2016.1166560

Kearl, Holly. 2010. *Stop Street Harassment: Making Public Places Safe and Welcoming for Women*. E-book. Santa Barbara: ABC-CLIO.

Kulshrestha, Juhi, *et al*. 2017. "Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417-432. New York: ACM. DOI: https://doi.org/10.1145/2998181.2998321

López-Paredes, Marco, & Branco Di Fátima. 2022. "Memética: la reinvención de las narrativas en el mundo digital, protestas sociales y discursos de odio." In *Visiones contemporáneas: narrativas, escenarios y ficciones,* edited by Oliver C. Márquez, Alicia Parras Parras, & Eva Hernández Martínez, 25-37. Madrid: Fragua.

Matsuda, Mari J. 1989. "Public Response to Racist Speech: Considering the Victim's Story." *Michigan Law Review* 87(8): 2320-2381. DOI: https://doi.org/10.2307/1289306

Milano, Stefano, Mariarosaria Taddeo, & Luciano Floridi. 2020. "Recommender Systems and Their Ethical Challenges." *AI & Society* 35(3): 957-967. DOI: https://doi.org/10.1007/s00146-020-00950-y

Miranda, Sofia, *et al*. 2022. "I love to hate! The racist hate speech in social media." *Proceedings of the 9th European Conference on Social* Media, 137-145. Krakow: Academic Conferences International (ACI). DOI: https://doi.org/10.34190/ecsm.9.1.311

Miró-Llinares, Fernando, & Jesús Javier Rodríguez-Sala. 2016. "Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy." *International Journal of Design & Nature and Ecodynamics* 11(3): 406-415. DOI: https://doi.org/10.2495/DNE-V11-N3-406-415

Moran, Mayo. 1994. "Talking about Hate Speech: A Rhetorical Analysis of American and Canadian Approaches to the Regulation of Hate Speech." *Wisconsin Law Review* 1994: 1425-1514.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Obermeyer, Ziad, *et al*. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366(6464): 447-453. DOI: https://doi.org/10.1126/science.aax2342

Olmos, Ana, *et al*. 2020. *Jóvenes, redes sociales virtuales y nuevas lógicas de funcionamiento del racismo: Etnografía virtual sobre representaciones y discursos de alteridad e identidad*. Madrid: Centro Reina Sofia sobre Adolescencia y Juventud. DOI: https://doi.org/10.5281/zenodo.3666178

Schwarzenegger, Christian, & Anna Wagner. 2018. "Can it be hate if it is fun? Discursive ensembles of hatred and laughter in extreme right satire on Facebook." *Studies in Communication and Media* 7(4): 473-498. DOI: https://doi.org/10.5771/2192-4007-2018-4-473

Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer." *The New York Times*, 10 March. Available at https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Valerio, Lizette Martínez. 2022. "Hate messages toward the LGTBIQ+ community in Instagram profiles of the Spanish press." *Revista Latina de Comunicación Social* 80: 364-388. DOI: https://doi.org/10.4185/RLCS-2022-1749

Van Dijck, José, Thomas Poell, & Martijn De Waal. 2018. *The Platform Society: Public Values in a Connective World*. New York: Oxford University Press.

Vesnic-Alujevic, Lucia. 2012. "Political participation and web 2.0 in Europe: A case study of Facebook." *Public Relations Review* 38(3): 466-470. DOI: https://doi.org/10.1016/j.pubrev.2012.01.010

Vilar-Lluch, Sara. 2023. "Understanding and appraising 'hate speech'." *Journal of Language Aggression and Conflict* 11(2): 279-306. DOI: https://doi.org/10.1075/jlac.00082.vil

Vincent, James. 2016. "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day." *The Verge*. Available at https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Vogels, Emily. 2021. *The State of Online Harassment*. Washington, D.C.: Pew Research Center. Available at https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Waldron, Jeremy. 2012. *The Harm in Hate Speech*. Cambridge: Harvard University Press.

Daniele Battista. Ph.D. in Social Theory, Digital Innovation, and Public Policies. Currently, he holds a research fellow position at the Department of Political and Social Studies at the University of Salerno. His primary areas of research interest revolve around the field of media theory, with a specific focus on the intricate relationship between social networks and political communication. He is member of editorial boards of several journals and member of the Italian Association of Sociology.

Jessica Camargo Molano. Assistant lecturer of Sociology of Electronic Arts and Sociology of Multimedia Entertainment at the University of Salerno, Italy. PhD student at the International Telematic University "UniNettuno". Her PhD research focuses on the algorithm, investigated both as a tool or creator of works through Artificial Intelligences, and as a support, certification of possession or even as true and own work of art in the case of NFT (Non-Fungible-Token). Currently, she is a visiting researcher at the University of Luxembourg.

Received on 25 August and accepted for publication on 31 October 2023.

How to cite this article

[Chicago Style]
Battista, Daniele, & Jessica Camargo Molano. 2023. "How AI Bots Have Reinforced Gender Bias in Hate Speech." *ex æquo* 48: 53-68. DOI: https://doi.org/10.22355/exaequo.2023.48.05

[APA Style – adapted]
Battista, Daniele, & Molano, Jessica Camargo (2023). How AI Bots Have Reinforced Gender Bias in Hate Speech. *ex æquo,* 48, 53-68. DOI: https://doi.org/10.22355/exaequo.2023.48.05