International Journal of Bullying Prevention (2022) 4:55–65
https://doi.org/10.1007/s42380-021-00105-7

ORIGINAL ARTICLE

AI Content Moderation, Racism and (de)Coloniality

Eugenia Siapera[1]

Accepted: 24 August 2021 / Published online: 6 September 2021

Abstract
The article develops a critical approach to AI in content moderation adopting a decolonial perspective. In particular, the article asks: to what extent does the current AI moderation system of platforms address racist hate speech and discrimination? Based on a critical reading of publicly available materials and publications on AI in content moderation, we argue that racialised people have no significant input in the definitions and decision making processes on racist hate speech and are also exploited as their unpaid labour is used to clean up platforms and to train AI systems. The disregard of the knowledge and experiences of racialised people and the expropriation of their labour with no compensation reproduce rather than eradicate racism. In theoretically making sense of this, we draw influences from Anibal Quijano's theory of the coloniality of power and the centrality of race, concluding that in its current iteration, AI in content moderation is a technology in the service of coloniality. Finally, the article develops a sketch for a decolonial approach to AI in content moderation, which aims to centre the voices of racialised communities and to reorient content moderation towards repairing, educating and sustaining communities.

Keywords AI · Hate speech · Racism · Content moderation · Decolonial theory · Digital platforms

Introduction: Race and Content Moderation

Content moderation on social media has proven to be a very complex matter. From something of an afterthought, it has emerged as a defining attribute of platforms (Gillespie, 2018). As social media acquire more users, and as more users upload contents, the importance of content moderation is likely to increase. Additionally, given the significant implications of problematic contents, states and international organisations are becoming increasingly concerned with developing an appropriate regulatory framework that platforms will then be asked to apply. Digital hate speech is associated with economic hardship, chilling effects and mental health issues. Specifically, targets of hate speech face economic hardship because they are excluded from a source of potential income, as Jane (2018) has shown.[1] Secondly, online hate speech silences its targets, or in other words it prevents them from exercising their rights (Penney, 2020) and therefore unfairly discriminates against them. Thirdly, the mental health issues suffered by targeted group members have been conclusively shown in past research (Anderson, 2013). These negative implications of digital hate speech exacerbate the inequalities and discrimination faced by certain communities, and in particular by racialised communities. As Lentin (2020) has shown, race still matters, because it is used to justify and sustain asymmetrical power relationships, in which racialised communities are exploited and their opportunities significantly diminished. Moreover, instead of the post-racial claims that race is no longer relevant (Goldberg, 2015), the opposite seems to be the case; as Mbembe (2017) has argued, the logic of race as a system of subjugation, exploitation and dispossession is expanding.

Because of this persistence and expansion of racial logics, it becomes important to identify and discuss the mechanisms and techniques by which these are sustained. This article is

✉ Eugenia Siapera
eugenia.siapera@ucd.ie

[1] School of Information and Communication Studies, University College Dublin, Stillorgan Road, Belfield, Dublin 4, Ireland

[1] Jane (2018) argued that cyberhate involves professional and economic harms to people, especially women. For example, Citron (2014) found that employers did not invite for interview women who have been subjected to image-based abuse. In other cases, content creators and those using social media professionally, such as YouTubers, 'influencers', journalists or politicians are subjected to systematic online hate and abuse, have reported that they have been reluctant to create or post more contents or interact online, therefore suffering economic hardship to the extent that their income and professional reputation is linked to their online presence. An investigation by the BBC showed that women are systematically targeted on the platform

looking to contribute to this discussion through focusing on the issue of racist hate speech in digital media.

In particular, the article focuses on the methods used to control racist hate speech in digital media, collectively known as content moderation (Gillespie, 2018), examining the role of Artificial Intelligence (AI) systems developed to address racist hate speech on digital platforms. The parameters within which content moderation operates are set by the reality of the content volume generated by users, the business and operational needs of platforms, and the regulatory and policy approaches of states and international bodies (Gillespie, 2018). As a result of these pressures, but also as part and parcel of their identity as technology companies, digital platforms allocate considerable resources and energy in developing automated systems that track and downgrade or remove contents deemed problematic.

In this context, focusing on racialised people specifically, the article presents a conceptual discussion of the possibilities and problems associated with the use of AI in content moderation. The central question posed is: to what extent does the current AI moderation system of platforms address racist hate speech and discrimination? Based on a critical reading of publicly available materials and publications on AI in content moderation, we argue that racialised people are reduced to either passive recipients of AI moderation with no significant input in the decision-making processes on racist hate speech or to low paid and even unpaid workers trying to clean up the platforms from contents that harm them. This removal of agency results in further oppression and exploitation, and in this manner reproduces rather than eradicates racism. In theoretically making sense of this, we draw influences from Anibal Quijano's (2000, 2007a, b) theory of the coloniality of power and the centrality of race, concluding that in its current iteration AI in content moderation is a technology at the service of coloniality. What may be an alternative position? In the final section of the article, following Quijano (2007a, b) and more recent works by Benjamin (2019), Birhane and Guest (2020), Mohamed et al. (2020) and Adams (2021), we develop a sketch for a decolonial approach to AI in content moderation.

In outlining these arguments, the article is divided into four sections. The first section discusses the theoretical framework of coloniality. The second section provides a description and discussion of content moderation and contextualises the shift to AI, based on a critical reading of materials published by platforms, posts and transcripts by senior executives and media reports. The third section develops a critique of AI in content moderation from a decolonial perspective. The final section outlines a decolonial approach to AI in content moderation.

Race and the Coloniality of Power

The term coloniality of power was suggested by Anibal Quijano (2007a) to explain and theorise the new forms of power at a distance that emerged in a post-colonial world. While countries that were former colonies achieved national independence, Quijano argued that they were and still are under the control of Western countries. This is because the structures of power and hegemony that were set up in the colonial era remain in place. This is especially the case with the notion of whiteness and the racialisation of subjectivity. Quijano's perspective is particularly valuable in the present context because it is centring the construct of race and because it understands it as a shifting technique of power across the world system.

Quijano (2007a) analysed these structures of power as they are deployed across four areas: control of economy, through practices such as land appropriation and labour exploitation; control of authority, such as institutions and governance processes; control of gender and sexuality, including family and education; and control of subjectivity and knowledge, including the processes of knowledge creation and formation of subjectivity. Quijano traces the historical emergence of these systems of control in the years that followed the European expansion and conquest of the Americas. These systems and power structures in place did not end when the direct colonial rule by Europeans was dismantled. Rather, they remain in forms that Quijano terms the coloniality of power (see also Quijano, 2000).

These power structures and systems of control revolve around the construction of race. Race becomes the organising principle of not only the colonial system, but the whole world (Quijano, 2007b). As Quijano (2007b: 45) put it, race is constituted as "the basic criterion for social classification of the entire world's population", so that humanity is organised around racialised identities, with white European identity at the top and every other identity arranged below. This arrangement was and is used to build and justify a system of domination, which includes a distribution of labour and flow of trade, and in general a "Eurocentering of capitalist world power" (Quijano, 2007b: 46). Race is therefore an outcome of colonial domination that has become pervasive in every aspect of global capitalist power. Racism, which then is the expression of this racial organisation of the world system, is an omnipresent reminder of this system of power and domination. Because it is omnipresent it constitutes one of the principal arenas of struggle and conflict (Quijano, 2007b).

The coloniality of power, which is Quijano's shorthand term for all these historically located relations of power

Footnote 1 (continued)

OnlyFans, often using racial slurs, leading some of them to close their accounts and therefore losing income (Croxford, 2021).

revolving around the construct of race, is evident in the organisation of natural resources, labour, the sphere of institutions, and the domain of knowledge. The expropriation of the labour and resources of racialised subjects, the development of institutions of governance from which racialised subjects are either excluded or limited to only certain roles, and the dismissal of any forms of knowledge except those associated with the techno-scientific rationality produced by Europeans are some of the main aspects of coloniality of power.

Quijano's approach makes clear the embeddedness of historically determined relations of power in the world system. Race is therefore not a biological category and racism is not an attitude or behaviour carried by individuals; rather they are both outcomes of the specific social organisation that emerged in early modernity along with European expansionism. The two most crucial insights offered by Quijano are firstly that race and racism are not epiphenomenal but fundamental and foundational aspects of modern societies; and secondly that they must be understood as dynamic and shifting, with dimensions that span across and structure the world system (see also Mbembe, 2017).

New technologies do not therefore emerge outside these systems; rather, they are part and parcel of current formations of economic, political and social power. In other words, race and racism are already embedded in technologies and the task for critical researchers is to understand their involvement in the reproduction of dominant structures of power (c.f. Feenberg, 1991). The next section will discuss and contextualise content moderation, in order to subsequently formulate a critique arguing that content moderation systems reproduce existing power structures and act as a technology of coloniality.

Content Moderation

A crucial aspect of digital platforms argues Tarleton Gillespie (2018) is that they all must moderate while disavowing moderation. Platforms that rely on users and their information must take steps to ensure that information deemed problematic is identified and controlled. This section begins with a brief discussion of the definition and current understandings of hate speech, the origins of content moderation, and the role of AI. This discussion relies on a critical reading of publicly available materials, including posts, speeches, and interviews given by senior platform executives, media articles and transparency reports. We identify the following issues: a failure to identify racism as a structure of power and address it as such; a lack of any substantial contribution by racialised people; and the labour conditions of human moderators. We then find that AI systems used in content moderation repeat what human moderators are doing at scale, while also detracting from focusing on the problematic definitions of racist hate speech and the politics of race and racism.

Defining Hate Speech

In the days of 'move fast and break things' — Mark Zuckerberg's famous quote for Facebook that has since become defunct — digital platforms rolled out their products with little concern for their potential impact. In these early days, Facebook did not have a moderation policy, other than excluding pornography, and did not even have a reporting mechanism for users to report inappropriate contents (Cartes, in Viejo-Otero, 2021). When such a functionality was built, both Facebook and Twitter, adopted an 'operational' approach to moderation policy, which emerged from the contents that users were reporting. Teams of in-house moderators would meet and discuss the reports, determine which should be removed or not, and then use these decisions to draft and update their relevant policies (Viejo-Otero, 2021). This dynamic approach to content policy, driven by what users reported, eventually led to the formulation of a set of clearly stated community guidelines (YouTube[2]) or community standards (Facebook[3]). These were also informed by existing legal frameworks, especially laws governing illegal contents.

The approach to hate speech is very similar across the platforms of Facebook, YouTube and Twitter. The policies on hate speech are loosely based on the main legal international instruments, conventions and declarations. The two most directly relevant include the tri-partite International Bill of Human Rights, comprised of the Universal Declaration of Human Rights (UDHR, 1948), the International Covenant of Civil and Political Rights (ICCPR, 1966), and the International Covenant of Economic, Social and Cultural Rights (ICESCR, 1966); and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD, 1965).[4]

The current platform policies on hate speech reflect the main points of these instruments and revolve around the notion of protected characteristics. Facebook, for instance, refers to race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease as protected characteristics. YouTube also

[2] The Guidelines are found here: https://www.youtube.com/intl/ALL_ie/howyoutubeworks/policies/community-guidelines/

[3] Found here: https://www.facebook.com/communitystandards/

[4] For a detailed discussion of these see Siapera et al. (2018).

includes age, immigration status, victims of a major violent event and their kin and veteran status. Content that incites to violence, discrimination or hatred against people who fall under these categories is considered hate speech and will be removed.

These definitions provide the framework within which platforms develop relevant policies. In general, neither the platforms, nor other institutions, for example the European Union, wish to expand, change or otherwise modify the definitions. This is clear in the Code of Conduct, which is a voluntary instrument signed by all the major platforms and the European Commission. The Code of Conduct has four main stipulations: (i) timely removal of illegal hate speech; (ii) hiring more content moderators; (iii) working with civil society in the removal of illegal hate speech (iv) and implementing a monitoring process to check for compliance. The recent Digital Services Act (European Commission, 2020) and the Audio-Visual Media Services Directive (AVMSD, 2018) that seek to regulate platforms do not provide any further definitions, relying on the existing EC Framework Decision of 2008 for illegal hate speech.

The definition of what constitutes hate speech therefore remains surprisingly unchanged since the mid-1900s (Siapera et al., 2017). In platforms, race is listed among many other protected characteristics, with no attention paid to the specificity of the experience of racism and racial exploitation. Indeed, the complete absence of any reference to the concrete historical circumstances that gave rise to race and racism as discussed by Quijano and many others, is striking. There is, for example, no distinction between races, so that contents that are against black people are equivalent to contents that are against white people. There is therefore no understanding or appreciation of the differences between the oppressed and the oppressors (Siapera & Viejo-Otero, 2021). The direct outcome of this race-blind approach in the definition of hate speech is that by equating oppressors and oppressed, it ends up privileging the former.[5] In other words, if the definition of what constitutes hate speech cannot in principle make a distinction between racist contents and criticisms of whiteness, it is inevitable that whiteness is protected and the system of power it represents remains intact. Additionally, because there is no focus on sharpening the definition, all the emphasis by the platforms (and the European Commission) is placed on the enforcement of policies for assessing and potentially removing contents, and the effectiveness of relevant measures. And it is at this level that the question of moderation enters the picture.

Human Moderation

The moderation of content on platforms is a key activity, as Gillespie (2018) has argued, because it effectively creates and maintains a product that is then sold to users. It is part of platforms' brand identity, which they then use to attract and maintain users and engagement. The decisions, for example, by both Facebook and YouTube to exclude sexually explicit/pornographic contents or graphic violence created platforms meant to have a wide appeal. Content moderation is therefore an essential and structural feature of platforms. This began as human labour, as teams of human moderators considered contents that were flagged as potentially containing hate speech or other categories of problematic contents. This human moderation, in turn, raised three important issues concerning: (i) potentially subjective decisions lacking standardisation; (ii) the labour practices and working conditions for moderators; (iii) the psychological costs of continuous exposure to hateful contents.

When in 2017 *The Guardian* and other major newspapers leaked the materials used by Facebook to moderate contents, journalists noted the complexity of rules and the inconsistencies in their policy. An example used by *The Guardian*, is the distinction between credible and non-credible threats: "Remarks such as "Someone shoot Trump" should be deleted, because as the then head of state he was in a protected category. But it can be permissible to say: "To snap a bitch's neck, make sure to apply all your pressure to the middle of her throat", or "fuck off and die" because they are not regarded as credible threats." (Hopkins, 2017, non-paginated). Moderators then must make decisions on contents based on rules that are difficult to interpret and apply and that may be modified often to include more examples of clarifications. A decision that a piece of content constitutes hate speech may have to rely on a content moderator's specific interpretation of the content and the policy. Additionally, Facebook and other platforms' policies are valid globally, regardless of cultural context (Gillespie, 2018). This means that moderators in different countries will have to understand and apply these policies in the same way, making any standardisation problematic.

These difficulties are further compounded by the labour conditions of moderators who are almost always externally contracted workers. They have certain targets to meet, i.e., contents that they must review, within a certain period. In *The Guardian* materials, it is estimated that moderators often have around 20 s to decide (Hopkins, 2017). Roberts (2019) estimates that there are some 100,000 content moderators across the world, and that they are in low paid, precarious or short-term contracts, working for companies that

[5] As Lentin (2007, 2020) and Goldberg (2015) have pointed out, race blind or post racial approaches refuse to acknowledge the historical specificity of racism against those deemed 'not white' and racialised, the privileges that have accrued to certain groups of people because of racism, and the operation of racism as a form of systemic and institutional discrimination. The focus on the relative lack of obvious, 'frozen' racism to prove that racism is no longer a problem, but rather a matter of a few 'bad apples'. In this manner, such approaches leave the current status quo largely unchanged and do not touch the privileges that the history of race and racism have bestowed on those constructed as white.

provide services to the platforms rather than for the platforms themselves. Their work is to review content that has been flagged as inappropriate and decide as soon as possible to move to the next piece of content. The job is of low status, and moderators are not expected to have any skills other than adequate knowledge of the language of the contents they moderate. A quick search of content moderator adverts on LinkedIn shows that this is an entry level position. For example, an advert for a Dublin-based content moderator in Arabic lists as mandatory criteria a 'high school diploma or equivalent' and at least a B2 level of Arabic, and while 'affinity and cultural awareness of political and social situations regarding the relevant market' are desirable, they are not required. These labour conditions as Roberts (2019) observes are exploitative. While content moderators constitute the unseen labour that is required to sustain the platforms, they enjoy none of the prestige or benefits that Facebook, Google or Twitter employees have. Roberts (2019: 70) observes that content moderators engage in "digital piecework" and are not offered any protections and benefits; for example, US content moderators were not given any health insurance in contrast to those directly employed by the platforms.

Roberts (2019) further exposed the toll of dealing with toxic contents on the mental health of content moderators. As one of her informants put it: "I can't imagine anyone who does [this] job and is able to just walk out at the end of their shift and just be done. You dwell on it whether you want it or not" (Max Breen, quoted in Roberts, 2019: 115). Confronted with violent, explicit, gross, hateful contents hour after hour and day after day has been linked to post-traumatic stress disorder (PTSD). In 2018 Facebook paid $52 m in compensation to almost 10,000 content moderators in the US who had suffered PTSD (BBC, 2020). While platforms and the companies that manage content moderation on their behalf (for example, Accenture, CPL/Covalen, Cognizant) provide 'wellness supports', content moderators feel that their mental health is not adequately protected, and that often companies evade any responsibility for this through non-disclosure agreements and even getting workers to sign waivers (RTE, 2021).

Yet despite these problems that have received extensive publicity, platforms are still bound to moderate their contents. It is in the context of public scrutiny in the media and pressure by international bodies, such as the European Commission, that the shift to AI occurred.

AI Based Automated Moderation

While all platforms use automated content moderation, Facebook is the one that has been more transparent about it and has signalled very clearly the intention to use artificial intelligence. Mark Zuckerberg did this himself in the Blueprint for Content Governance and Enforcement (Zuckerberg, 2018). In this and other documents, Zuckerberg makes the case that accuracy and consistency are two key issues for content governance; that the scale of content posted on Facebook cannot be dealt with exclusively by human labour; and that repetitive tasks are better performed by computers (Zuckerberg, 2018; Newton, 2019). Additionally, AI systems can provide detailed metrics on content that was 'actioned', i.e., content that was flagged and led to a decision, and this is an important parameter that fulfils the demands for transparency and efficiency of content moderation. Finally, as digital platforms operate within the ideology and value system of Silicon Valley, providing technologically innovative solutions is their preferred route (c.f. Barbrook & Cameron, 1996; Morozov, 2013). In short, AI provides a "desirable, inevitable, unavoidable" (Gillespie, 2020) solution to the problems posed by content moderation and the human labour involved.

Elaborating on these points, firstly, a platform such as Facebook hosts billions of pieces[6] of content per day, making content moderation at this scale a task that cannot be dealt by humans alone. According to Zuckerberg (2018: non-paginated), because of advances in Artificial Intelligence but also "because of the multi-billion-dollar annual investments we can now fund" it is possible to "identify and remove a much larger percent of the harmful content — and we can often remove it faster". In his 2018 testimony to the US Congress, Zuckerberg argued that "over the long term, building AI tools is going to be the scalable way to identify and root out most of this harmful content" (cited in Harwell, 2018). In the same testimony he referred to AI more than 30 times, alluding to the importance given to AI within Facebook. According to Joaquin Quiñonero Candela, Facebook's director of Applied Machine Learning, the increased importance of AI is indicated by the physical location of the FAIR (Fundamental AI Research) and AML (Applied Machine Learning) teams next to Mark Zuckerberg's own office, at Building 20, the main office at the Menlo Park headquarters (Hao, 2021).

Secondly, AI moderation is seen as a tool that will help protect moderators from the emotional and mental burden of viewing and acting on toxic contents. Specifically, Mike Schroepfer, Facebook's Chief Technology Officer, explained that AI tools can "get the appropriate decisions on the content without having the same sort of emotional impact on the person viewing it. So there's a ton of work that I can't represent in 30 s here, but it is a key focus for all the tools teams to sort of reduce dramatically the human impact it would have by looking at this terrible stuff" (Schroepfer, cited in Newton, 2019). More recently, representatives for content moderators published an open

[6] According to Hootsuite, in January 2021 users post 1.55 pieces of content daily, and Facebook has 2.74 billion active users.

letter saying that "management told moderators that we should no longer see certain varieties of toxic content coming up in the review tool from which we work — such as graphic violence or child abuse, for example" (cited in Foxglove, 2020).

But how are these systems deployed and what do they do? Facebook began using AI systems for proactive moderation — that is, for picking up potentially problematic contents without users flagging them — in 2016. Facebook's Fundamental AI Research (FAIR) developed and refined their own in-house systems, such as Deep Text and FastText (an open-sourced library for text representation and classification). Their systems can be used for text and images, as well as for text on images and videos, while they have also developed models for different languages. Their system XLM-R (RoBERTa) works across 15 languages; their new system Linformer, introduced in late 2020, is a more efficient and precise classifier while another system (Reinforcement Integrity Optimizer or RIO) has been developed for optimizing hate speech classifiers that automatically review all content that gets uploaded on Facebook and Instagram (Schroepfer, 2021).

In a typical application of AI, Facebook deploys a model to predict whether something is hate speech based on the extent of its similarity with contents previously identified as having violated existing policies; then, another system determines the action to be taken, for example, to delete it, demote it, or send it for human review. Their newest system RIO has improved these processes by training the classification systems based on the performance not only of the prediction (how accurately it detected hate speech) but also on how successful the enforcement was (for example, how many people were protected from seeing the content) (Facebook, 2020). While RIO improved the overall system performance and efficiency of the training data, Linformer enabled training to consider contextual features, and XLM-R to consider additional languages (Schroepfer, 2021). Facebook's transparency reports indicate that together these systems have an impressive rate of success. In the final quarter of 2020, 97.1% of the hate speech contents that were removed were proactively detected by Facebook's AI systems, before they were reported and before they were seen by any users (Facebook Transparency Report, 2021). Overall, Facebook estimates that the prevalence of hate speech on the platform dropped from an estimated 0.11% in the third quarter of 2020 to 0.07% in the fourth quarter (this means that out of 10,000 pieces of content 11 and 7 were estimated to contain hate speech for the two quarters respectively). Facebook attributes this drop mainly to improvements in proactive moderation through the introduction of the systems discussed above.

While Facebook reports astonishing successes, other platforms have mixed results. Because of Covid-related restrictions and the shift from working from home, content moderator teams were furloughed, and their number decreased. This led YouTube to rely more on AI systems in the last few months. According to a report by the *Financial Times*, the systems removed proactively 11 million videos in the second quarter of 2020, twice the usual rate (Barker & Murphy, 2020). The accuracy of the removals was also lower, as about 50% of the removal appeals were upheld when AI was responsible for the removal, compared with less than 25% of those upheld when decisions were made by human moderators.

While researchers have questioned the extent to which AI can consider the context and nuances of language (Caplan, 2018), the response of platforms such as Facebook has typically been that the technology is constantly improving and that in combination with human moderation these systems will eventually be highly effective in recognizing hate speech (Schroepfer, 2021). However, AI for content moderation has been criticised not only in terms of its accuracy, but on a more conceptual basis. Gillespie (2020) argues that platforms are involved in a discursive justification of AI in content moderation in a way that becomes self-fulfilling and meets the platforms' own ambitions for further growth: "platforms have reached a scale where only AI solutions seem viable; AI solutions allow platforms to grow further." Gorwa et al. (2020) argue that AI introduces further opacity into moderation decisions, because the ways in which AI algorithms work is neither clear not accountable. Additionally, Gorwa et al. argue that AI is presumed to be un-biased, but this in fact obscures the ways in which certain viewpoints are privileged; they note that classifiers operate based on certain formulations of toxic, racist or misogynist content, and therefore ignoring others. A third critical point proposed by Gorwa et al. (2020) concerns the de-politicisation of the politics of content moderation; in proffering AI as an answer to problems of content moderation, platforms position themselves as invisible infrastructures and hide the political decision making behind the types of contents deemed acceptable or not. These are much more fundamental problems that cannot be addressed via technological improvements. Such arguments open a space for a critique of the role of AI in content moderation and its relationship with existing power systems. However, they tend to be general points that do not consider the effects of the shift to AI for race and racism more specifically. The next section develops a critique based on decolonial values.

A Decolonial Critique of AI for Content Moderation

Following up on the point on the depoliticization of content moderation raised by Gorwa et al. (2020) and considering theories of coloniality and race as a key structure of power, it is evident that the shift to AI is obscuring the

political questions of how racist hate speech is codified in content moderation policies, techniques and technologies. As Quijano (2007b) argued, racism is the omnipresent reminder of the racial constitution of the world. This is supported by a specific division of labour, in which racialised subjects' labour is undervalued and by a dismissal of the knowledge and experiences of racialised subjects. As we will argue below, these are evident in the ways in which platforms conceptualise hate speech, in how they disregard the voices of the communities targeted by racism, and they exploit the labour of racialised subjects. In doing so, platforms and the AI systems they develop and deploy for racist hate speech constitute colonial technologies and materialise racial inequalities. These arguments will be supported by looking at how the policies against hate speech obscure the question of race and racism, and silence those who are targeted. In this context, AI moderation is merely implementing these policies at scale. At the same time, in training AI systems to classify racist hate speech, platforms rely on the mostly unacknowledged and undervalued work of racialised people.

Despite the successes claimed by platforms, the question remains: to what extent do AI systems resolve the problems of moderating racist hate speech on platforms? Mark Zuckerberg (2018) suggests that these "are not problems you fix, but issues where you continually improve. Just as a free society will always have crime and our expectation of government is not to eliminate all crime but to effectively manage and reduce it, our community will also always face its share of abuse. Our job is to keep the misuse low, consistently improve over time, and stay ahead of new threats." This discursive construction coming from Facebook's CEO is important, because as Haupt (2021) has argued, the way in which Facebook presents itself to the world offers an insight into its position as a powerful social actor. In this respect, it is important to see that Facebook sets for itself the goal of managing and reducing hate speech rather than contributing to the eradication of racism. The policy rationale for the removal of hate speech in the Community Standards makes no reference to racism or discrimination: "We believe that people use their voice and connect more freely when they don't feel attacked based on who they are. That is why we don't allow hate speech on Facebook.[7]" In these terms, the relevant policies of Facebook are oriented towards safeguarding Facebook's main values of voice and connectivity (Viejo-Otero, 2021) without necessarily addressing racism and discrimination. Both human and AI-driven moderation implement these policies and participate in the management, but not eradication, of hate speech.

[7] Available at: https://transparency.fb.com/policies/community-standards/hate-speech/

The 'management of hate' proposed by Facebook in practice allows certain types of hate speech to circulate. A key issue when it comes to digital hate speech and racism is that the illegal part is sustained by a base of 'low-level' 'banal' racist contents which do not qualify as hate speech, but nevertheless scaffold and build up towards the kind of serious, violent, dehumanising racist speech that is illegal in Europe (Sharma, 2017; Siapera, 2019). The 'management' of this, as Facebook puts it, involves removing some of the top part of the pyramid, the overt instances of illegal hate speech, while allowing the 'banal' kind, occasionally demoting it.

In these terms, a fundamental problem with this emphasis on AI moderation, is that it diverts attention from the definitions and the overall conceptual framework that Facebook, and other platforms has built around hate speech and its implications. It hides the question of who participates in the definitional and enforcement frameworks and decisions. While there are various critiques and discussions of the definitional context (Matamoros-Fernandez, 2017; Siapera & Viejo Otero, 2021), the main argument here is that the definitions of (racist) hate speech do not consult or consider the views, experiences and socio-cultural understandings of key targeted groups. As we saw earlier, content moderators are only required to be adequate speakers of the language in which they moderate and are tasked with reports on all kinds of contents, from spam to extreme violence, including hate speech; they therefore lack specialized knowledge about racism, and related hateful discourses. AI systems in turn get trained on data on hate speech, which are based on judgements made by moderators. Schroepfer (in Kahn, 2020) reports that Facebook's classifiers compare a new piece of content with those that have already been removed following human moderation. Similarly, Gillespie (2020) observes that human moderators produce training data so that they are later replaced by AI. So even if we accept an increased capacity for efficient and standardised classifications and decisions, AI systems in practice just apply moderators' judgement at scale. These judgements are based on a fit between the content and the descriptions in the Community Standards/Terms of Service. While the Code of Conduct stipulates the involvement of civil society organisations, some representing targeted groups, their role is to promote counternarratives and/or to act as 'trusted flaggers'. The 'trusted flagger' role is to flag or report hateful content that has come to their attention through a direct, fast track system which ensures a quick decision by the platforms. Given the vast amounts of content posted on platforms everyday, this is a marginal role with limited impact. Those directly targeted by hate speech are still excluded or marginalised from participating in the definition of what constitutes hate speech. This exclusion turns users likely to be targeted by hate speech into passive observers or recipients of decisions made by automated systems into which they have no conscious input, no

oversight and no control. The only available role for them is that of a trusted flagger, who is there to report contents that violate content policies, but without having a say as to how these policies are constructed in the first place.

The shift to AI in content moderation has not therefore introduced an opening up or fundamental rethinking of the policies on racist hate speech. Rather, it involves the implementation of these policies on a larger scale, enabled by the automation of the process of reviewing contents. In doing so, AI is continuing the silencing of racialised groups that are targeted by hate speech, who still make no contribution in defining racist speech.

As we have seen, the only avenue that racialised communities have for providing input to content moderation systems is through either individually flagging or reporting racist contents, or through occupying the position of trusted flaggers. This is unpaid work, that community members who are targeted by racist speech undertake on behalf of their community. These reports, in turn, end up in datasets used for training AI systems to recognize when violations have occurred or not. In this manner, the unpaid labour of targeted communities and their allies ends up on the one hand cleaning up Facebook of individual instances of racist speech, and on the other training algorithms that can then do this at scale. Since, as we saw above, Facebook is looking to remove hate speech to allow more people to post contents, which is the basis of their revenue model, this unpaid labour returns profit for the platform. For the members of targeted communities whose labour is expropriated with no compensation, this is exploitative and discriminatory as it disproportionately affects racialised communities.

The silencing of members of racialised communities, the disregard of their knowledge and experiences, is, as we saw in the discussion of decolonial theory, a key part of the coloniality of power. In dismissing the views and knowledge of targeted communities, and in excluding them from participating in the definition of their own oppression, content moderation policies continue and reinforce the colonial tactic of dismissing the knowledge of colonized communities. Additionally, content moderation systems are extracting and expropriating the labour of these communities because they use it to sustain the platforms without any compensation for the communities. Extraction and expropriation are also key elements of the coloniality of power as Quijano (2007a) has shown. Taken together, the silencing of the voices of racialised communities and the extraction of their labour materialize racial inequalities, since these communities are treated unequally, and the process of content moderation affects them disproportionately. In this manner, content moderation reinforces whiteness, understood as the current system of racial power, along with the exclusions and privileges it affords (Harris, 1993). While these tactics are part of how content moderation is conceived, the use of AI is built upon and sustains the current system of moderation, while also detracting from issues of definition by constructing the moderation of hate speech as a problem of scale. It is because of these aspects that we understand AI in content moderation as a colonial technique: it enables the governing of racialised subjects from a distance and through excluding them from meaningful input and in this manner supports the existing racial division of power.

Returning to the original question posed in the beginning of the article, concerning the extent to which AI in content moderation addresses the challenges of racist hate speech and discrimination, we have found that not only does it not address these, but it ends up reinforcing and reproducing existing racial power structures, constituting a technique of coloniality. Can this be challenged? The next section presents a discussion of a possible alternative.

AI and Content Moderation: A Decolonial Approach

As we have argued, AI systems do not emerge in a vacuum but are already part and parcel of existing relations of power. This has been identified by research into the development and application of these systems. For instance, Ruha Benjamin (2019) has shown how the application of AI has resulted in the deeper embedding of racial codes that already permeate society. To use an example, both predictive policing and algorithms predicting recidivism have been trained on historical data and make use of demographic data, post codes, insurance rates and so on. While these systems and their designers assume that these data are neutral and represent reality in an accurate manner, in fact, they re-encode the systemic inequality that is already found in the criminal justice system. It has therefore been shown that such algorithms classify black people as more likely to offend or reoffend (Angwin et al., 2016). The use of these systems in policing, housing, risk assessments, health and so on evidently works in ways that reinforce the systemic oppression of racialised communities. Benjamin therefore urges for a reconceptualizing technologies as abolitionist tools, by which she means tools designed to address oppression and discrimination and to help reimagine a new, more equitable world. Using specific examples of AI tools, such as 'Appolition', an app that collects bail money and redirects resources to local organisations supporting communities, Benjamin explains how technologies can be reimagined in ways that entail freedom rather than control. In a similar vein, Birhane and Guest (2020) argue against the tokenistic appointment of a few black men and women in positions of influence without changing and challenging the underlying structures that uphold current systems of power. They therefore argue for a gradual process of decolonization through looking at the history of the disciplines involved in the

production of new technologies such as AI, and an opening up of the field to those who challenge established orthodoxies (Birhane & Guest, 2020). While Birhane and Guest propose a historically nuanced understanding, Mohamed et al. (2020) look to the future, proposing the development and integration of foresight in the design of technological systems, such that it can recognize and preempt any systemic discrimination. The common thread of these views is that AI systems can ultimately be disengaged from the history of coloniality and made to serve the purpose of dismantling it. For this to be accomplished, however, it is not enough to adopt an ethics or fairness approach as this can reenact colonial practices, such as the assumption of a universal ethics, which is in fact European/Western (Adams, 2021). Adams (2021) argues that the problem with current AI ethics approaches is that they draw decolonial thinking into existing discourses of ethics rather than using it in the way it was intended: as "an invocation to make intelligible, to critique, and to seek to undo the logics and politics of race and coloniality that continue to operate in technologies and imaginaries associated with AI in ways that exclude, delimit, and degrade other ways of knowing, living, and being" (Adams, 2021: 190).

In order therefore to develop a decolonial approach to AI in content moderation, we must think not of the classification techniques themselves but of the broader context within which content moderation of racist hate speech and AI are both embedded. This, as we discussed, is a context in which racist hate speech, its moderation on platforms and the AI systems developed to serve this moderation are all involved in perpetuating systems of oppression. A decolonial approach therefore requires a clear contribution to ending the exploitation and domination that have emerged from the histories of colonialism and racial capitalism. It follows that AI for content moderation should be developed within a broader engagement with these legacies and therefore outside of platforms, as these have a self-serving agenda that is not concerned with questions of racism and racial justice. As such, AI in content moderation should be seen as part of an arsenal of an approach that seeks to dismantle systems of domination, which include the ways in which speech, discourses, and ideologies circulate and are used to justify current race/gender hierarchies and subjugation of other forms of knowing and being. From this point of view, AI techniques may be useful for knowing the ways in which race and racism operate currently than for removing or demoting content based on a limited and ultimately problematic definition of hate speech. Ultimately, rather than embedding these technologies in content moderation systems whose goal is to remove or demote contents, they should be used to repair, heal, educate and sustain communities (c.f. Adams, 2021).

In such a perspective, AI in content moderation is developed as part of an approach that encodes and operationalises values of equality, justice and liberation/emancipation and not as a pseudo-technosolution. From a decolonial point of view, these are the values that should guide the development of AI systems rather than a race-blind version of equality. This view echoes Benjamin's (2019) call for the development of technologies as abolitionist tools.

In practice, these values require that oppressed people are given voice and their experiences valued. Definitions of racist speech should therefore emerge from those who are at the receiving end and not from those systems or processes involved in the oppressing. To avoid exploitation, AI classifiers should be trained on data produced by people whose consent and input is sought and fully compensated for. Finally, these tools should remain in the control of those who developed them and used in the interests of those oppressed and discriminated.

Conclusion

The article sought to develop a critical approach to AI technologies used in the moderation of racist hate speech. In doing so, it assumed a decolonial perspective, drawing upon the work of Anibal Quijano (2007a, b) who historicised race as emerging out of the colonial expansion of Europeans, and used as part of a system of power that justified the domination of those designated as non-white. Quijano argues that although former colonies have become formally independent, coloniality still structures the world in an unequal manner. Decoloniality is therefore the process by which the current systems of power, revolving around race and the racialisation of people, their continued silencing and exploitation, can be questioned and dismantled.

With this in mind, we turned to the issue of racist hate speech and its treatment in platforms and their content moderation systems. We saw those current definitions of hate speech do not include the views or positions of those targeted by racist hate speech; and secondly, they assume a procedural and general approach, disregarding the specific history of racism, making no references to historically oppressed groups. Platforms then imported the spirit of these definitions and used it in their terms of service and content policies. In applying these policies, platforms have used human moderation. This is low skilled, repetitive labour mainly undertaken by outsourced workers in short term contracts. AI based moderation emerged as a more efficient solution to the massive increase of problematic content and to the intense public scrutiny of platform content policies. AI moderation is addressing questions of scale, accuracy and standardisation, while also protecting users from being exposed to hate speech and other problematic contents by removing them proactively.

Technological advances and investment in AI systems point to its increased importance for platforms. However,

AI systems still apply the same generic definitions and in the final analysis, they perform the job of relatively unskilled human moderators at scale. Moreover, in the development of AI systems to classify, remove or demote hate speech contents, platforms such as Facebook use training data that come from contents previously reported by those attacked or their allies and found to violate content policies. This labour is unacknowledged and uncompensated, and therefore exploitative. Facebook has further made clear that its goal is to manage hate speech to enable more people to post contents rather than address questions of racism and discrimination against specific racialised communities targeted by hate speech.

On this basis, the critique of AI in content moderation revolves around the following key issues: the continued silencing of racialised communities, the discounting of their experiences, and the exploitative use of their free labour. Because these aspects affect disproportionately racialised communities, the present content moderation systems of platforms and the ways in which they conceive and develop AI tools are reinforcing the current racial status quo and can therefore be characterised as colonial technologies.

A decolonial perspective to AI in content moderation should therefore contribute to the dismantling of current racial logics, designed and developed with this goal and following the decolonial values of equality, justice and emancipation. The centring of the voices of the oppressed in describing their own oppression is key to such a perspective, as is the acknowledgment and compensation of their labour. Rather than the individuating aim of removing discrete pieces of content, moderation technologies should aim to repair, educate and sustain communities.

To conclude, we found that in its current iteration AI in content moderation represents and serves to perpetuate systems of domination. If the current iteration of AI systems is seen as the go-to technology to save us from racist hate speech, it is bound to fail. In the context of racist hate speech, the question and challenge for us is to ask if we can create AI systems that embed and represent different values and serve the interests of equality, justice and emancipation.

Funding Open Access funding provided by the IReL Consortium.

Declarations

Conflict of Interest The author declares no competing interests.

References

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews, 46*(1–2), 176–197.

Anderson, K. (2013). Diagnosing discrimination: Stress from perceived racism and the mental and physical health effects. *Sociological Inquiry, 83*(1), 55–81.

Angwin, J., Larson, J., Mattu, S. & Kirchner L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barbrook, R., & Cameron, A. (1996). The Californian ideology. *Science as Culture, 6*(1), 44–72.

Barker A. & Murphy, H. (2020). YouTube reverts to human moderators in fight against misinformation, in *The Financial Times*, September 20, available at: https://www.ft.com/content/e54737c5-8488-4e66-b087-d1ad426ac9fa

BBC. (2020). Facebook to pay $52m to content moderators over PTSD, May 12, available at: https://www.bbc.com/news/technology-52642633

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity press.

Birhane, A. & Guest, O. (2020). Towards decolonising computational sciences. arXiv preprint arXiv:2009.14258.

Caplan, R. (2018). Content or Context Moderation? New York, NY: Data & Society Research Institute. Available at: https://datasociety.net/output/content-or-context-moderation/

Citron, D. K. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Croxford, R. (2021). Under the Skin of OnlyFans. BBC News, July 17, available at: https://www.bbc.com/news/uk-57269939

European Commission. (2018). Audiovisual Media Services Directive (AVMSD) available at : https://ec.europa.eu/digital-single-market/en/revision-audiovisual-media-services-directive-avmsd

European Commision. (2020). Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, available at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-parliament-and-council-single market-digitalservices-digital-services

Facebook. (2020). Training AI to detect hate speech in the real world, November 19, available at: https://ai.facebook.com/blog/training-ai-to-detect-hate-speech-in-the-real-world

Facebook. (2021). Facebook Transparency Report, available at: https://transparency.facebook.com/

Feenberg, A. (1991). *Critical theory of technology* New York: Oxford University Press.

Foxglove. (2020). Open letter from content moderators re: pandemic, available at: https://www.foxglove.org.uk/news/open-letter-from-content-moderators-re-pandemic

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2), 2053951720943234.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Goldberg, D. T. (2015). *Are we all postracial yet?* John Wiley & Sons.

Gorwa, R., Binns, R. & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 2053951719897945.

Hao, K. (2021). Artificial intelligence: How Facebook got addicted to spreading misinformation. In MIT Technology Review, March 11, available at: https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Harris, C.I. (1993). Whiteness as property. In *Harvard law review*, 1707–1.

Harwell, D. (2018). AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how. In The Washington Post, April 11th, available at: https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

Haupt, J. (2021). Facebook futures: Mark Zuckerberg's discursive construction of a better world. *New Media & Society, 23*(2), 237–257.

Hopkins, N. (2017). Revealed: Facebook's internal rulebook on sex, terrorism and violence, in *The Guardian*, May 21st, available at https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

ICCPR. (1966). International Covenant on Civil and Political Rights, available at: https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

ICERD. (1965). International Convention on the Elimination of All Forms of Racial Discrimination, available at: https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx

ICESCR. (1966). International Covenant on Economic, Social and Cultural Rights, available at: https://www.ohchr.org/EN/ProfessionalInterest/Pages/CESCR.aspx

Jane, E. A. (2018). Gendered cyberhate as workplace harassment and economic vandalism. *Feminist Media Studies, 18*(4), 575–591.

Kahn, J. (2020). Facebook's A.I. is getting better at finding malicious content—but it won't solve the company's problems, in *Fortune*, November 19, available at: https://fortune.com/2020/11/19/facebook-ai-content-problems-artificial-intelligence/

Lentin, A. (2007). *Racism*, OneWorld Publications.

Lentin, A. (2020). *Why race still matters*. John Wiley & Sons.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian Race-Based Controversy on Twitter. *Facebook and YouTube, in Information, Communication & Society, 20*(6), 930–946.

Mbembe, A. (2017). *Critique of black reason*. Duke University Press.

Mohamed, S., Png, M.T. & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*(4), 659–684.

Morozov, E. (2013). To save everything, click here: The folly of technological solutionism, in *Public Affairs*, March 5.

Newton, C. (2019). Transcript from a leaked discussion between Mark Zuckerberg and Facebook employees, Oct 1, available at: https://www.getrevue.co/profile/caseynewton/issues/the-world-responds-to-the-zuckerberg-leak-also-more-leaking-202477?utm_campaign=Issue&utm_content=view_in_browser&utm_medium=email&utm_source=The+Interface

Penney, J. (2020). Online Abuse, Chilling Effects, and Human Rights, in Dubois, E. and Martin-Bariteau, F. (eds.), Citizenship in a Connected Canada: A Research and Policy Agenda, Ottawa, ON: University of Ottawa Press, Available at SSRN: https://ssrn.com/abstract=3620520

Quijano, A. (2000). Coloniality of power and Eurocentrism in Latin America. *International Sociology, 15*(2), 215–232.

Quijano, A. (2007a). Coloniality and modernity/rationality. *Cultural Studies, 21*(2–3), 168–178.

Quijano, A. (2007b). Questioning "race". *Socialism and Democracy, 21*(1), 45–53.

Roberts, S.T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

RTE, (2021). 'It took me ages to sleep': Facebook moderators on the things they can't unsee, January 29, available at: https://www.rte.ie/news/primetime/2021/0128/1193745-facebook-moderators-mental-health-leo-varadkar/

Schroepfer, M. (2021). Update on Our Progress on AI and Hate Speech Detection, Facebook, February 11, available at: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/

Sharma, S. (2017). Theorizing online racism: The stream, affect & power laws. In *AoIR conference* (pp. 19–21).

Siapera, E. (2019). Organised and ambient digital racism: Multidirectional flows in the Irish digital sphere in *Open Library of. Humanities, 5*(1), 1–34.

Siapera, E., & Viejo-Otero, P. (2021). Governing hate: Facebook and digital racism. *Television & New Media, 22*(2), 112–130.

Siapera, E., Moreo, E. & Zhou, J. (2018). *Hate track: Tracking and monitoring online racist speech*. Irish Human Rights and Equality Commission. Available at: https://www.ihrec.ie/app/uploads/2018/11/HateTrack-Tracking-and-Monitoring-Racist-Hate-Speech-Online.pdf

Siapera, E., Viejo-Otero, P. & Moreo, E. (2017). Hate speech: Genealogies, tensions and contentions. In *Association of Internet Researchers (AoIR) conference* (pp. 19–21).

UDHR, (1948). Universal Declaration of Human Rights, available at: https://www.un.org/en/universal-declaration-human-rights/

Viejo-Otero, P. (2021). The Governance of Hate Speech, Unpublished PhD Thesis, DCU.

Zuckerberg, M. (2018). "A Blueprint for Content Governance and Enforcement." https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/