RACISM IN THE DIGITAL AGE: THE IMPACT OF SOCIAL MEDIA ALGORITHMS ON PUBLIC DISCOURSE

ÁNGELES SOLANES CORELLA[*]
NACHO HERNÁNDEZ MORENO[**]

Abstract: This paper explores the intersection of social media and racism, focusing on how algorithms and their biases perpetuate racial inequalities in the digital public sphere. By examining the evolution of the public sphere, the role of algorithms, and the spread of racist narratives, the study highlights the ethical implications for democracy and digital governance. The findings underscore the need for regulatory interventions to mitigate algorithmic bias and ensure fair public discourse, promoting a more inclusive and equitable digital environment.

Keywords: Algorithmic Bias, Social Media, Racism, Public Discourse, Digital Ethics.

1. Introduction

The intersection of social media[1] and racism is a critical area of study, particularly in understanding how algorithmic bias perpetuates racial inequalities in the digital public sphere. Terms like "algorithm" and "artificial intelligence" (AI) have become commonplace in everyday language; however, their impact on individuals and society is not yet addressed as a priority in the public or political debate. Although seemingly neutral, algorithms are designed to achieve a specific outcome through a series of calculable steps and formalized instructions. They have also become narrative devices in the media, used to describe the

[*] Professor of Philosophy of Law and Politics at the Universitat de València. Member of the Human Rights Institute of the Universitat de València. E-mail: angeles.solanes@uv.es. ORCID: https://orcid.org/0000-0002-4207-6113

[**] Assistant Professor (*Profesor Ayudante Doctor*) of Philosophy of Law and Politics at the Universitat de València. Member of the Human Rights Institute of the Universitat de València. E-mail: nacho.hernandez@uv.es. ORCID: https://orcid.org/0000-0002-6686-662X

[1] The term *social media* has been preferred over *social networking sites* due to its broader scope, which encompasses the latter (Aichner *et al*. 2021, p. 215; Carr & Hayes, 2015, p. 49). *Social networking sites* are understood as a subset of social media primarily focused on interaction and relationship-building (e.g., X, Facebook, LinkedIn), whereas *social media* is more generally associated with platforms that promote content creation (e.g., YouTube, TikTok) (Solanes & Hernández, 2024, p. 109). However, given that *social media* functions as an umbrella term encompassing both kinds of platforms and that the issue explored in this paper relates to each of them, we consider it the most appropriate expression to capture the digital context addressed in this study.

This article was written as part of the R&D&I project PID2023-147896OB-I00, «Derechos humanos e inmigración: repensar la integración en tiempos oscuros», funded by the Ministry of Science, Innovation and Universities

magical processes behind friend suggestions on social networks or product and service recommendations tailored to users' preferences in cyberspace (Cabrera, 2021).

Regarding artificial intelligence, the absence of a universally accepted definition allows Floridi to frame AI more as agency than intelligence. According to this author, the issue is not whether a machine is intelligent or capable of thinking, but rather that its behavior may appear to be the product of human decision-making (Floridi, 2024, p. 72). As such, both a procedure (related to algorithms) and a behavior or agency (concerning AI) may seem harmless at first glance. However, their practical application conceals potential biases that can foster discrimination and racism —both directly within the digital realm and indirectly in physical spaces.

This study focuses precisely on this issue: investigating how algorithms in social media contribute to the spread and normalization of racist narratives. For the purposes of this study, public discourse is understood as the exchange of ideas and information within the public sphere, where individuals and groups discuss matters of common interest (Habermas, 1986). To effectively address societal issues, it must be governed by the principles of orderly argument, mutual respect, and sincerity (Hannon, 2023, p. 337). The central research question guiding this work is: How do algorithms on social media contribute to the spread and normalization of racist narratives? The working hypothesis is that algorithmic bias on these platforms amplifies racist content, thereby shaping public discourse in ways that reinforce racial inequalities.

The relevance of this research lies in the growing influence of social media on public opinion and the increasing concern over algorithmic bias. As digital platforms become central to political, social, and cultural interactions, it is crucial to understand how they shape public discourse. This study addresses the ethical implications of algorithmic bias, emphasizing the urgent need for regulatory interventions to ensure fair and inclusive public dialogue. By exploring the intersection of technology and racism, the research contributes to ongoing debates on digital ethics, democracy, and social justice. Rather than addressing this issue through a more comprehensive or detailed regulation of the circumstances that enable the rise of racism in cyberspace, platforms invoke *freedom of speech* as a shield to avoid monitoring the content hosted within their spaces. In doing so, they also evade their responsibility as sovereign entities of a virtual world where discrimination, as well as incitement to violence, hatred, and hostility against racialized otherness, has found refuge.

The paper is structured to provide a comprehensive analysis of the interplay between social media, algorithmic bias, and racism. It begins by examining how the internet and social media have transformed the public sphere, focusing on the role of digital platforms in shaping public discourse. This is followed by an analysis of the business models of social media, emphasizing their profit-driven nature and the consequences for public dialogue. The paper then traces the evolution of racism, moving from biological determinism to culturalism, and explores how it is perpetuated through algorithms. Furthermore, it discusses various types of algorithmic bias and their impact on racial inequalities, analyzing how social media platforms facilitate the spread and normalization of racist content. The conclusion summarizes the findings, addresses the broader implications for democracy, digital ethics, and policymaking, and offers recommendations for mitigating algorithmic bias to foster fair and inclusive public discourse.

2. The Evolution of the Public Sphere in the Digital Age

The Internet has revolutionized our lives and is now an essential tool in most of our daily activities, whether related to work, study, leisure, consumption, or interaction with other individuals and institutions. Social media has followed a similar trajectory, becoming fundamental to understanding the current political, social, and cultural landscape of our societies. This phenomenon exemplifies how a tool originating in cyberspace can exert a significant impact on the physical and tangible world, influencing electoral outcomes while also having the potential to incite, promote, or foster hostility, violence, discrimination, or hatred against specific groups. Consequently, social media plays a crucial role in analyzing and investigating public discourse and the rise, normalization, and legitimization of new forms of contemporary racism.

Such is their relevance that their emergence and current ubiquity have contributed to what Habermas (2023) describes as the "platformization" of the public sphere—the metaphorical space in which public opinion is shaped, a key driver of media and political agendas. This concept refers to the social space where individuals engage in rational discourse, free from state interference, to deliberate on matters of common concern (Habermas, 1989, p. 136)[2]. While in its traditional sense it relied on print media, radio, and television to shape public discourse, cyberspace now serves as an alternative domain where digital platforms mediate political and social interactions. However, this shift has brought challenges that raise concerns about the integrity of democratic deliberation in the digital age.

Historically, the public sphere existed in physical spaces like salons and coffeehouses, where structured debates shaped collective consciousness. With capitalism's expansion, the bourgeoisie recognized the political implications of private commercial

[2] Habermas's theory of public discourse centers on the idea that democratic legitimacy arises from rational-critical debate among citizens in the public sphere. However, in his evolutionary analysis of the public sphere, the German philosopher also warns that the emancipatory potential of this space has been undermined over time (Habermas, 1986). What once functioned as a counterpower has gradually been colonized by private interests, such as the media, political marketing, and economic actors, which transform public discourse into a mechanism for managing consent rather than fostering genuine democratic will-formation. As a result, public opinion is no longer the product of free and rational deliberation among citizens, but is increasingly mediated and shaped by the strategic interests of a minority. This problem is further exacerbated in the digital age. With the advent of the internet and especially social media, the public sphere has become platformized—migrating from physical spaces into virtual environments controlled by large private corporations (Habermas, 2023). Although these digital platforms present themselves as open arenas for public discourse, they are primarily governed by algorithms and business models designed to maximize engagement and profit. This evolution raises significant concerns about the distortion of deliberative processes, as virality, emotional appeal, and algorithmic curation often displace reasoned debate, further distancing the public sphere from the ideals of communicative rationality.

interests, prompting a counterpublic sphere that used the press to challenge institutions and demand transparency. However, these debates remained limited to educated elites—jurists, doctors, priests, and academics—whose intellectual authority often outweighed social hierarchy (Habermas, 1986; 1989). The printing press revolutionized public discourse, spreading information and fostering shared consciousness. Whereas newspapers initially challenged power, they gradually became instruments of persuasion and propaganda, prioritizing advertising over objectivity, creating a manipulative publicity that eroded democratic deliberation.

Habermas (1986) warns that media commercialization fosters an irrational public sphere dominated by strategic messaging rather than genuine public deliberation. Public opinion, as a legitimizing force of power—even in authoritarian regimes (Arendt, 2006, pp. 56-57)—is shaped by the interaction between individuals, information flows, and media influence (Sartori, 1993, pp. 59-60). Ideally, public discourse serves as a check on institutional power, fostering a healthy democratic process. However, the public sphere, increasingly dominated by mass communication, has progressively distanced public opinion from genuine deliberation, making it more vulnerable to manipulation (Habermas, 1986; 1989).

The referred massification is seen in contemporary society and its current public sphere. The internet has introduced new mechanisms for public engagement, providing tools that allow individuals to express opinions and challenge institutional authority in real-time. The digitalization of communication enables viral dissemination of information, granting citizens greater oversight over political and economic actors. However, this same dynamic exposes the public sphere to manipulation by powerful interests, including state and non-state actors who exploit digital platforms for ideological or commercial gain. An effective public sphere must be universally accessible, facilitate the exchange of perspectives, and promote collective deliberation (Seeliger & Sevignani, 2022, p. 8). Although social media and, in particular social media, seemingly provide a virtual space open to anyone who registers and logs in, allowing users a certain degree of freedom to personalize their profiles, interact, and engage with digital tools to participate in shaping public discourse, the reality, as will be examined in the following sections, is far more complex and nuanced. These platforms, which are often perceived as spaces of freedom and empowerment, diverge significantly from this ideal in practice. As we will explore, the fact that social media are designed by private entities driven by economic interests, coupled with a business model reliant on the attention economy and algorithmic mediation, precludes them from functioning as genuinely neutral, public, or free spaces.

The impact of digital technologies on the public sphere has been both transformative and destabilizing, as the so-called "engineering of consent" has undermined the autonomous formation of public opinion, as traditional and new media have been successful in "getting people to support ideas and programs" (Bernays, 1947, p. 114). Manipulative strategies create the illusion of democratic participation while, in reality, reinforcing power asymmetries that benefit private interests (Habermas, 1986). This phenomenon is particularly evident in digital environments, where social media platforms mediate political discourse through algorithmic amplification. Algorithms are fundamental to the functioning of essential services and infrastructures in information societies. While their potential to enhance individual and collective well-being is undeniable, so too are their associated risks, as algorithms are not ethically neutral (Floridi, 2024, p. 205). Although users may perceive themselves as active participants in shaping public opinion, the dynamics of digital discourse are often dictated by a small number of influential actors or automated bots. The algorithms governing these platforms prioritize engagement metrics over deliberative quality, amplifying content designed to maximize user retention rather than fostering rational debate.

The platformization of the public sphere, as described by Habermas (2023), marks a fundamental shift from traditional media structures to digital ecosystems characterized by decentralized content production. Unlike the press, which historically mediated public discourse through editorial oversight, the internet enables individuals to engage in public communication with minimal barriers to entry. While this democratization of discourse provides opportunities for marginalized voices to be heard, it also raises significant concerns regarding the spread of misinformation and the erosion of journalistic standards.

Seeliger and Sevignani (2022) identify three defining characteristics of the platformized public sphere: digitalization, commodification, and globalization. Digitalization facilitates bidirectional communication through networked platforms, while commodification reflects the economic exploitation of user-generated content. The globalization of digital communication further complicates regulatory efforts, as transnational information flows challenge traditional frameworks of political accountability. These factors collectively contribute to the fragmentation of public discourse, making it increasingly difficult to establish a shared deliberative space.

The internet's impact on democracy is thus paradoxical. While it has expanded the reach of public discourse, it has also introduced structural limitations that hinder rational deliberation. The abundance of communicative activity ironically restricts meaningful engagement, as the sheer volume of information undermines the ability to prioritize substantive discussions. This contrasts with traditional media, where editorial oversight provided a degree of coherence in public communication (Rodríguez-Izquierdo Serrano, 2017). The rise of "echo chambers" in digital spaces has further exacerbated ideological polarization, limiting exposure to diverse viewpoints (Miller *et al.*, 2021)[3]. Social media algorithms reinforce pre-existing biases (such as confirmation bias, wherein users gravitate towards information that suits their preexisting beliefs) by curating content that aligns with user preferences, reducing the likelihood of cross-cutting dialogue and making individuals more susceptible to misinformation (Barberá, 2020; Menczer & Hills, 2020, p. 59). This segmentation of public discourse undermines the deliberative ideal envisioned by Habermas, fostering insular communities rather than inclusive debate.

[3] "Eco chambers" (or "filter bubbles") is a widely used term in this context that can be defined as digital environments in which individuals are predominantly exposed to information, opinions, and perspectives that reinforce their preexisting beliefs, due to algorithmic filtering and selective engagement, thereby amplifying ideological polarization and reducing exposure to dissenting views.

Ultimately, the public sphere in its digital form reflects both the promises and perils of technological innovation. While the internet has facilitated unprecedented levels of political engagement, it has also subjected public discourse to the influence of private corporations that prioritize profit over democratic deliberation. Habermas (1986) warned of the risks posed by commercialized media, arguing that the manipulation of public opinion serves entrenched interests under the guise of public benefit. These concerns remain relevant in the digital age, where algorithmic governance and surveillance-based business models further entrench economic and political power asymmetries.

In this context, public opinion is increasingly shaped by users—who may be individuals or automated entities—rather than by citizens in the traditional sense. Whereas citizenship is tied to state recognition and the legal framework of nationality, user status is conferred through acceptance of the contractual terms imposed by social media corporations. Unlike citizenship, which remains an exclusive legal category, user status is broadly inclusive, requiring only access to an internet-enabled device. Paradoxically, this user identity aligns more closely with the universal human rights ideal than traditional citizenship, as digital platforms do not impose restrictions based on nationality. Nevertheless, engagement in public discourse is contingent upon participation as a user rather than as a citizen, as only users have access to the digital tools necessary to shape contemporary political narratives. However, social media platforms are not neutral arenas for public debate; they are commercial enterprises designed to maximize engagement and profit. While they may create the illusion of unrestricted freedom, they are fundamentally engineered to influence behavior, fostering prolonged engagement through mechanisms that subtly constrain individual agency.

As democratic systems grapple with these challenges, the future of the public sphere will depend on regulatory interventions that promote transparency, accountability, and inclusive participation. The digital public sphere must be restructured to ensure that it remains a space for genuine deliberation rather than a tool for ideological manipulation. Achieving this balance will require a concerted effort to reconcile the emancipatory potential of digital technologies with the need for ethical governance and institutional oversight.

3. Social Media: Seemingly Public and Neutral Places under the Control of Private Interests

The notion of neutrality in social media is fundamentally misleading. Although these platforms are widely marketed as open and accessible spaces where content is generated by users, this apparent neutrality is largely illusory. In practice, not all participants represent genuine individuals, and not every piece of content is the product of voluntary, unmediated expression. Rather, social media companies are primarily profit-driven enterprises whose core objective is to maximize user engagement. Their platforms are carefully engineered to stimulate continuous interaction, ensuring that every click, view, and share contributes to a business model focused on revenue generation. Consequently, user participation is not a spontaneous, neutral act but is instead cultivated within an ecosystem specifically designed to optimize engagement and, by extension, profit.

Beneath the surface of free-to-use services lies a sophisticated business model that capitalizes on massive data collection and targeted advertising. Social media harvest extensive amounts of information—ranging from basic demographics to intricate details of browsing behavior and interpersonal interactions. This data is then processed using advanced algorithms, which analyze patterns to deliver highly personalized advertising. For example, Facebook's advertising infrastructure enables marketers to target individuals based on a myriad of factors, including location, age, interests, and past online activity. Real-time bidding mechanisms further complicate this landscape; companies compete in digital auctions where the highest bidder secures prime placement in a user's newsfeed (Wang, Zhang, & Yuan, 2017). Such strategies are designed to ensure that advertisements are not only visible but are also tailored to elicit a high degree of user engagement, thereby reinforcing the commercial interests of the platform at the expense of genuine user autonomy.

The inherent business model is predicated on the principle of maximizing user engagement, a goal that fuels an endless cycle of data collection and content personalization. The longer an individual remains on a platform, the more behavioral data is captured, enabling further refinement of targeted advertising and content curation. This feedback loop—where extended user interaction leads to even greater personalization—can be understood as a core component of the so-called attention economy. In this paradigm, every second a user spends on a platform is commodified; algorithms are deliberately designed to exploit psychological vulnerabilities and keep users engaged for as long as possible. Features such as autoplay, infinite scrolling, and push notifications are not incidental but are central to this strategy. They manipulate cognitive processes by providing intermittent rewards that encourage prolonged, often compulsive, usage patterns, maximizing screen time, which in turn increases ad exposure and revenue generation. Social validation mechanisms, such as likes, shares, and comments, further intensify this effect by tapping into innate human desires for approval and social recognition (Bhargava & Velasquez, 2021, p. 327).

The deliberate design of these platforms creates conditions analogous to addiction, though unlike chemical substances, this dependency is engineered through digital interfaces and algorithmic cues. Over time, excessive engagement with social media can impair cognitive functions—reducing users' ability to concentrate, make sound decisions, and engage in meaningful real-world interactions (Ophir, Nass & Wagner, 2009). The removal of natural stopping points, such as the traditional cues that signal the end of a conversation or the closing of a book, further exacerbates these effects. In the context of infinite scrolling and constant content updates, users are deprived of the cognitive breaks necessary to regulate their consumption. This engineered compulsivity not only diminishes individual autonomy but also fosters an environment in which misinformation and low-quality content can thrive. As platforms prioritize what keeps users engaged, sensationalist and emotionally charged content is often amplified at the expense of nuanced, balanced discourse (Menczer & Hills, 2020). Algorithmic amplification thereby not only sustains user attention but also reshapes public discourse by privileging content that aligns with commercial imperatives over factual accuracy or societal benefits (Gillespie, 2018, pp. 194-196; Aridor *et al*., 2023, p. 9; Metzler & Garcia, 2023, p. 736).

One of the most concerning consequences of such algorithmic curation is the emergence of “filter bubbles” or so-called “echo chambers” previously referred to in the previous section. As users are repeatedly exposed to content that mirrors their pre-existing beliefs and preferences, their exposure to diverse perspectives diminishes dramatically. The platformization of the public sphere leads to a form of ideological insulation where individuals become confined to self-reinforcing loops of information. This phenomenon poses a direct threat to democratic pluralism because it prevents the kind of cross-cutting engagement necessary for a healthy public debate. Empirical studies have shown that algorithmic personalization contributes significantly to political polarization and the rapid dissemination of misinformation (Cinelli *et al*., 2021). Platforms like Facebook, YouTube, and TikTok employ reinforcement learning techniques that refine content recommendations based on previous user interactions. Consequently, as users increasingly engage with similar types of content, algorithms become adept at predicting and reinforcing these patterns, thereby narrowing the scope of available discourse and impeding opportunities for meaningful deliberation.

Moreover, the extensive reliance on algorithmic curation transforms the internet’s vast repository of knowledge into a mechanism of intellectual regression. This phenomenon has been aptly described as “mass self-communication,” a condition in which the overwhelming volume of personalized content prevents meaningful engagement with diverse perspectives. Rather than fostering enriched dialogue, the internet instead becomes a tool for reinforcing narrow, self-selected narratives—a state that some scholars have likened to “electronic autism” (Gonzálvez, 2011, p. 135). In such an environment, the foundational pillars of democratic civic engagement—public debate, shared social experiences, and exposure to differing viewpoints—are systematically weakened. Consequently, the collective construction of social reality is undermined, with profound implications for both individual autonomy and societal cohesion. Indeed, research indicates that users with more insular social media networks—those with fewer shared experiences and limited exposure to alternative perspectives—are more likely to respond aggressively to challenges that contextualize news or present opposing viewpoints (Kim, 2020, p. 524).

As Burke (2023) observes, maintaining an open mind within a closed system is inherently difficult, and questioning dominant structures becomes nearly impossible in the absence of alternatives. He highlights the role of ignorance in perpetuating ideological isolation and underscores the necessity of engaging with diverse perspectives to foster critical reflection. This is particularly relevant in the context of social media, where algorithmically curated content reinforces ideological echo chambers. Within these digital enclaves, individuals retreat behind a veil of ignorance—though not in the Rawlsian sense, which seeks to free individuals from personal biases to establish a fair and impartial social contract—but rather as a mechanism of self-imposed intellectual confinement. Shielded from dissenting views, these individuals find alternative perspectives increasingly disconcerting and difficult to confront over time.

Adding to the complexity is the concentration of control over online speech in the hands of a few dominant platforms. The centralization of information dissemination in these digital spaces raises profound concerns regarding the moderation and amplification of voices. Content moderation practices, which determine which voices are heard and which are silenced, play a crucial role in shaping public discourse. Just and Latzer (2016) argue that algorithms are not merely passive tools but active agents in constructing individual realities and shaping the broader social order. The personalization inherent in these algorithms fosters an extreme form of individualism, generating social fragmentation by reducing the occurrence of shared experiences that are vital for social cohesion. Therefore, individuals find themselves increasingly subject to the dictates of algorithmic control, resulting in a loss of both privacy and freedom (Just & Latzer, 2016, 254–255).

Beneath the facade of neutrality lies the illusion of free will. The pervasiveness and proactivity of algorithms, combined with users' limited understanding of their inner workings, weaken individual autonomy (Floridi, 2024). While social media platforms present themselves as neutral facilitators of communication and information exchange—implying that individuals bear full responsibility for their actions, interactions, and digital behavior—they simultaneously construct a curated perception of self-determination. This impression is reinforced through personalized feeds, customizable avatars, pseudonyms, and other filtering mechanisms. However, such customization does not reflect genuine user control but is ultimately dictated by commercial imperatives that shape the design and priorities of the digital landscape. Research on digital platform governance highlights the phenomenon of algorithmic nudging, whereby subtle design choices steer user behavior toward outcomes that serve platform interests (Susser, Roessler, & Nissenbaum, 2019). Even when platforms provide alternatives such as more transparent or chronologically ordered feeds, these options are frequently deprioritized or made less accessible, thereby constraining the possibility of authentic user agency (Bucher, 2018, pp. 149-150; Narayanan, 2023, pp. 40-41). This opacity in algorithmic decision-making not only diminishes digital self-governance but also raises significant ethical concerns regarding the extent to which individuals can exercise informed control over their online interactions.

The design of these algorithms is inherently biased. Rather than being neutral conduits for information, they are imbued with the commercial objectives of the platforms that deploy them. For instance, TikTok's "black box" algorithms, which remain largely opaque to users, are designed to curate content in a manner that subtly shapes perceptions and reinforces specific worldviews (Metzler & Garcia, 2023, p. 742). Moreover, the concentration of power in a handful of tech giants exacerbates concerns about the centralization of digital discourse. When content moderation policies are determined by a small number of corporate entities, the resulting control over what information is disseminated or suppressed can have far-reaching implications for democratic participation. The selective amplification of certain voices—often those that conform to profitable or politically expedient narratives—further underscores the need for robust regulatory oversight to safeguard the democratic integrity of digital public spheres[4].

[4] Article 34 of the Digital Services Act requires very large online platforms and very large online search engines to "identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services." In fulfilling this obligation, they must consider, for example, how the design of their recommender systems and any other relevant algorithmic systems may contribute to or influence such systemic risks.

In this context, it is unsurprising that racism is on the rise as a consequence of social media emerging as a crucial instrument within the contemporary, platformized public sphere. Algorithms dictate what users see and interact with, doing so in a manner that, as previously mentioned, maximizes user engagement with the ultimate goal of deriving economic profit from their attention. To achieve this, users are exposed to sensationalist content designed to elicit strong reactions and encourage prolonged platform usage. Populist and extremist views are particularly effective in capturing attention and serve as powerful tools for user retention. The more individuals are exposed to such perspectives, the longer they are likely to remain engaged. Moreover, if these reductionist, essentialist, and manipulative narratives resonate with users' preexisting beliefs, they will reinforce stereotypes or myths, which, through mere repetition, may come to be perceived as true. Echo chambers function as resonance spaces that prevent exposure to alternative viewpoints.

Extremist discourse, including racist rhetoric, which in traditional media is more susceptible to censorship due to editorial and regulatory filters, faces far fewer obstacles to dissemination on social media. Racism finds fertile ground in these platforms, bolstered by an interpretation of freedom of expression as an (almost) absolute right within the U.S. legal framework (Solanes & Hernández, 2024, p. 158), where the headquarters of the major social media corporations are located. This legal context enables racism to thrive, as social media platforms provide an ideal mechanism for its mass dissemination, with algorithms acting as facilitators of its propagation. Racism has not been eradicated; it remains present and is increasingly sustained and amplified by social media.

4. Racism in the Age of Algorithm Governance

4.1. From Biological Racism to Culturalism

Racism has evolved into a complex and multi-faceted social phenomenon deeply ingrained in historical, political, and cultural structures. Historically, racism was rooted in biological determinism, a pseudo-scientific belief that human races were hierarchically structured based on physical attributes. Over time, however, this form of racism has given way to a more insidious manifestation—cultural racism—where racial distinctions are increasingly framed in terms of cultural incompatibility rather than biological inferiority. This shift is not merely a change in terminology but reflects a broader transformation in the mechanisms of exclusion, where ideas of race continue to shape social hierarchies, albeit under different justifications.

Racism can be understood as a system of domination and subordination between groups, rooted in the racialization of differences. This system operates across interpersonal, institutional, and cultural dimensions, manifesting through various forms of invisibilization, stigmatization, discrimination, exclusion, exploitation, aggression, and dispossession (Aguilar Idáñez & Buraschi, 2016, p. 34). In their conceptual framework, these same authors (2019, p. 157) highlight two fundamental aspects of racism: first, its nature as a structured system of oppression rooted in asymmetrical power relations and a presumption of superiority; second, the process by which differences—whether physical, cultural, or social—are racialized and perceived as intrinsic and immutable. This perspective aligns with Wistrich's (1999, p. 2) view of racism as a form of heterophobia—a hostility toward the "other" that seeks to convert real or imagined differences into fixed, hereditary, and eternal attributes. These constructed distinctions serve to legitimize the marginalization of those deemed foreign, inferior, or primitive. Crucially, such social categorizations gain power when they are naturalized, appearing not as social constructs but as immutable realities.

Nonetheless, while race is not a natural concept nor a biological reality—and is scientifically debunked, morally reprehensible, socially unjust, and dangerous—it is not a spontaneous idea emerging from human perception but a historical, cultural, social, and ideological construct used to establish differences between groups (Aguilar Idáñez & Buraschi, 2019). According to Thomas's theorem, simply believing in the existence of this construct makes the consequences of such belief tangible. While races do not exist, racists do—individuals who believe in races that can be biologically defined (whether by phenotype or genotype), but also those who believe in their own superiority over others, even defending the rights of the "inferior race," or those who perceive multiple equal races.

Racism was once defined by its biological determinism and its inherently unequal nature (Blum, 2002, p. 4), and while the first characteristic may now be largely overcome, racism has not disappeared; it has evolved, now presenting itself with a moralizing democratic face (Aguilar Idáñez & Buraschi, 2019, p. 160). Whether in its classical or contemporary form, the core of racism remains the same: it asserts the dominant group's right to lead, control, subjugate, and exploit others (Wistrich, 1999, p. 2). To function as a system of oppression, racism requires social indifference, selective empathy, and the rejection of others, including the forcibly displaced (Aguilar Idáñez & Buraschi, 2019, p. 165). Racism has adapted by moving away from its biological foundations and instead creating new categories and hierarchies that align with changing social structures and power dynamics, ensuring its persistence in contemporary society.

Such adaptation to contemporary societies has created new forms of hierarchical classification that involve domination, subordination, and racialization of group differences that manifest in practices such as exclusion, criminalization, violence, expulsion, segregation, discrimination, and exploitation. These practices are justified within democratic frameworks, often invoking security and freedom as key principles. The pursuit of these two values drives the hierarchical categorization of groups, resulting in discriminatory treatment of the oppressed in favor of the dominant social group. This, in turn, uses the defense of its freedom and security to justify the establishment of a social and institutional structure that maintains the status quo of privilege, which is viewed as a fundamental right tied to moral superiority. What matters in this "democratic racism" is not the truth of the discourse, but its capacity to effectively manage power relations (Aguilar Idáñez & Buraschi, 2019, p. 164).

Ethnocentrism plays a key role in this new form of racism, where culture becomes the central element in distancing between groups and enables the establishment of a publicly acceptable social hierarchy based on the idea of morally superior cultures. Ethnocentrism, defined by Bizumic and Duckitt (2012) as a construct emphasizing the centrality and importance of the in-group, contributes to this perspective. Cultural supremacy, or the belief that one's culture is superior to others, is rooted in ethnocentrism but is further fueled by racist ideologies. Authors such as Bratt (2022, p. 207) argue that modern European racism manifests through the belief in the superiority of European culture, and the insistence on preserving the supposed purity of one's culture clearly reflects a racist perspective. In fact, Banton (1996) noted that opposition to immigration in European states during the previous three decades was more related to cultural differences than to supposed biological distinctions.

As well as ethnocentrism, xenophobia and identitarianism contribute to a racist climate by emphasizing cultural distinctiveness, which fuels exclusionary practices, reinforces subordination, and deepens the racialization of differences. As a result, contemporary racism gains greater acceptance than its more overt, traditional forms. Identitarianism has become a significant force advocating for the rejection of the "other," particularly migrants and refugees. This ideology has gained traction in many European states, bolstered by the rise of political parties that propagate exclusionary policies. The central tenet of identitarianism is the assertion of a homogeneous racial, cultural, religious, and linguistic identity that must be preserved at all costs. However, such an identity is a myth (Hernández, 2022, p. 42). Identity, as Balibar (2005) suggests, is inherently ambiguous and cannot be reduced to static, immutable attributes. Moreover, the claim to find fixed and absolute identity cores, entirely self-contained and unchanging, defined by stable and consistent characteristics that shape entities as immutable is erroneous (De Lucas, 2003). Nonetheless, as the Thomas theorem posits, if people believe in the reality of racial or cultural purity, the consequences of such beliefs become real, manifesting in policies of exclusion, discrimination, and violence against migrants and racialized groups.

Xenophobia is commonly understood as the fear or aversion toward foreigners. While this phenomenon is often described as a natural human response, research challenges this assumption, arguing that xenophobia is a social construct shaped by hierarchical identity frameworks. Thus, it is erroneous to consider that xenophobic rejection is inherent to human nature. Many cultures foster social harmony rather than the creation of hierarchies, and difference is not always seen as a threat. Furthermore, not all identities are defined by exclusion or by what they are not. The idea that xenophobia is a natural and inevitable response is therefore a myth, based on a particular perspective that frames identities in terms of hierarchical organization. Saito (2021) argues that this view serves as a useful tool for those in power: while racism is officially rejected in theory, discriminatory measures based on a person's foreignness—rather than their racial or ethnic identity—are met with greater acceptance.

This facilitates the implementation of exclusionary policies under the guise of national security or economic stability, even when they disproportionately target racialized groups. Racism and xenophobia intersect through their shared construction of the "other" as a threat. The difference between the two lies in their justifications: whereas xenophobia is centered on perceived foreignness, racism is supported by a belief in the inherent inferiority of the racialized other (Banton, 1996; Sundstrom & Kim, 2014, p. 35). Consequently, racism results in systemic oppression, including discrimination, exploitation, and social exclusion. Xenophobic attitudes often serve as a gateway to racist policies, as they create an environment where discrimination is justified based on perceived cultural incompatibility rather than explicit racial inferiority.

The rejection of migrants and racialized groups is justified through a modern form of racism that has moved beyond pseudo-biological justifications to embrace cultural differentiation. This shift reflects a broader trend where racism is embedded within discourses of security, national identity, and social cohesion. In multicultural societies, perceived cultural threats exacerbate tensions, leading dominant groups to reinforce exclusionary practices to protect their purportedly superior cultural identity. This results in the stigmatization of migrants and ethnic minorities, who are often associated with danger and criminality (Solanes, 2018, p. 60).

Wieviorka (1994, p. 42) explains how racism operates through two intertwined logics: inequality and difference. By reducing racism to social inequality (e.g., emphasizing only economic or class disparities) its distinct racial dimension is obscured, and it is treated as merely one aspect of broader social issues. Also, by framing racism only as cultural differentiation (e.g., seeing it as just a matter of cultural diversity) its structural effects are overlooked, reducing it to a benign difference in identity. What the author suggests is that racism emerges when these two logics—inequality and difference—are combined. Specifically, racism unifies these elements by assigning difference (e.g., cultural or ethnic distinctiveness) to a vulnerable minority group, which is then placed in a subordinate position, making it easier to justify discrimination and exclusion. Other characteristics of contemporary racism noted by Aguilar Idáñez and Buraschi (2019, pp. 161-162) are its own denial, its reduction to politically correct stances, color blindness or the lack of consideration of race's influence on power relations, reversed racism by minorities, national identity, binary polarization as a radical difference between groups, the victimization of the dominant group, or the blaming of the victim. The denial of racism, coupled with the reconfiguration of racist ideologies into seemingly democratic discourses, allows exclusionary practices to persist under the guise of cultural protectionism.

Furthermore, in addition to the traditional vehicles that have historically fostered racism—whether in its original or culturalist form—such as the family, society, the state, or mass media, the internet and social media have now joined this dynamic. Thanks to their unique characteristics, they have enabled the dissemination of ideas at a much greater speed and with a broader influence than these traditional instruments. Racism is on the rise.

4.2. Bias in the Machine: How Can Algorithms Be Racist?

How can seemingly neutral elements, such as algorithms or artificial intelligence, contribute to the rise of racist narratives in public discourse? The business model under study, as well as the role of these tools in monetizing users' attention and time on social media platforms, is closely related to this issue. This section will examine their specific impact on the perpetuation of racism in society, offering examples of various types of biases present in algorithmic and AI systems that contribute to this problem.

Algorithms, which are often perceived as neutral, are, in fact, not neutral and can inadvertently reflect and reinforce societal biases. As Floridi (2024, p. 226) points out, characteristics such as gender, religion, or race cannot simply be removed from training data to prevent both direct and indirect discrimination. This is because seemingly neutral variables, such as postal codes, can serve as proxies for inferring these characteristics, including race. Biases can emerge at various stages of data processing, from data collection to decision-making, and can result from factors such as historical human biases or incomplete or unrepresentative data used to train the algorithm (Lee, Resnick, & Barton, 2019). This may result in unintentional discrimination, one that can be more difficult to identify and address (Barocas & Selbst, 2016, p. 674).

Mehrabi *et al*. (2021) identify three primary sources of bias: data-driven, algorithm-driven, and user-driven bias, all of which may contribute to the reproduction of racial inequalities on digital platforms. The first category refers to biases embedded in the data used by machine learning training algorithms, which may result in skewed algorithmic outcomes. The second involves the introduction of bias through the algorithm itself, particularly as it modulates user behavior. The third encompasses biases inherent to users, which may be reflected in the information they generate and which is subsequently used to train machine learning models (Mehrabi *et al*., 2021, pp. 4, 7–8).

In the following paragraphs, the framework proposed by Mehrabi *et al*. (2021) will be connected with the typology developed by Silva and Kenney (2018), such that the specific biases identified by the latter are grouped within the three general categories delineated by the former. Accordingly, among the data-driven biases identified by Silva and Kenney (2018) are *training data bias* and *algorithmic focus bias*. Within the category of algorithm-driven bias, we find *algorithmic processing bias*. Finally, the user-driven biases include *consumer bias* and *feedback loop bias*. Silva and Kenney (2018) also describe other forms of bias—namely *transfer context bias*, *interpretation bias*, *non-transparency of outcomes bias*, and *automation bias*—which, however, do not neatly fit into any of the aforementioned categories. These forms of bias typically arise from user or individual action after the algorithm has produced an output, but they are not necessarily incorporated into the data used to train machine learning models. Each of these will be discussed in the following sections, as they all play a role in reinforcing racial inequalities on digital platforms, demonstrating how algorithms often perpetuate discrimination rather than mitigate it.

Training data bias occurs when the datasets used to train machine learning algorithms are filled with historical discrimination. For instance, if a social media algorithm is trained on data that overrepresents white users while underrepresenting racial minorities, it may result in the disproportionate amplification of voices from the former group. The absence of racialized people at various stages of programming and designing algorithms leads to racist outcomes and studies suggest that diversity decreases the likelihood of bias (Garcia, 2016, p. 114). Studies have demonstrated that image recognition software often misidentifies racial and ethnic minorities due to biased training datasets (Perkowitz, 2021), and in the social media context, this can manifest as the prioritization of content from white users, overshadowing racial minorities and diminishing discussions on racism.

Another form of bias, algorithmic focus bias, arises when an algorithm places greater emphasis on certain types of content while sidelining others. Social media platforms, designed to maximize user engagement, tend to favor polarizing but non-confrontational content. This can suppress important conversations about racial injustice, particularly as platforms are concerned with avoiding backlash or the withdrawal of advertisers. For instance, reports have shown that Meta's content moderation algorithms have disproportionately flagged posts addressing racism and suspended racialized users more often by that system (Lee *et al*., 2024).

Algorithmic processing bias refers to the way in which algorithms weigh variables to make decisions. If a social media algorithm heavily prioritizes "user engagement," it may give precedence to posts that resonate with the majority demographic or create eco chambers than increase the racial segregation of networks, operating as a sort of glass ceiling for minorities (Stoica *et al*., 2018; 2020). This can lead to the deprioritization of content that challenges white-centric perspectives on race, effectively sidelining discussions on systemic racism and further entrenching existing biases.

Transfer context bias arises when algorithmic outcomes are applied in a different context from the one originally intended, potentially leading to biased or discriminatory decisions. For instance, the use of credit scores in employee hiring processes is problematic because it assumes that a poor credit score correlates with poor job performance, despite the lack of conclusive evidence supporting such an association. This bias disproportionately affects groups that have historically faced restricted access to credit, thereby perpetuating structural inequalities. Within the context of social media, this type of bias can influence content moderation and the visibility of posts. For example, if an automated content review system classifies certain words or images associated with ethnic minorities as "sensitive" or "controversial" based on their usage in unrelated contexts, this may lead to the unjustified censorship of content from these communities. Consequently, transfer context bias can contribute to systemic racism by limiting the representation of specific groups and reinforcing dominant narratives.

Similarly, interpretation bias occurs when users' perceptions or actions introduce bias into the algorithmic process, that is, when they subjectively interpret ambiguous algorithmic outputs based on their own internalized biases. Even if an algorithm processes content impartially, user interpretation can influence its outcomes. Within the context of social media, this bias can contribute to racial discrimination in various ways. For instance, content moderation relies on automated flagging systems, but human moderators make final decisions. Due to interpretation bias, content created by racial minorities may be classified as "offensive" or "harmful" more frequently than similar content from other groups. Also, algorithmic filtering can reduce the visibility of minority voices if prejudiced users mass-report or downvote their content, leading platforms to misinterpret this activity as a quality issue, ultimately suppressing diverse perspectives.

Non-transparency of outcomes bias emerges when the rationale behind algorithmic decisions is opaque, even to the developers of such systems. Due to the complexity of machine learning and big data models, developers are often unable to explain why an algorithm arrived at a particular decision. This bias is particularly problematic in social media, where algorithms determine which content is displayed to users. If a system automatically decides that certain profiles or posts should have limited reach without providing an explanation, content creators from marginalized communities may experience reduced visibility without any means of appeal or adjustment. Furthermore, as many companies protect their algorithmic structures for commercial reasons, the lack of transparency prevents the detection and rectification of racial biases in content moderation and advertising.

Automation bias is another critical form, wherein users assume that algorithmic decisions are neutral and objective. Users may believe that newsfeeds curated by algorithms offer an unbiased representation of reality. However, these algorithms often prioritize mainstream narratives shaped by historically dominant racial groups, thus reinforcing whitewashed depictions of history and current events (Noble, 2018, p. 71). This assumption of neutrality, when in fact the algorithmic process is shaped by underlying biases, can contribute to the marginalization of minority perspectives.

Consumer bias is another dimension of algorithmic bias, emerging from the platform's adaptation to user preferences. If users engage more with content that downplays racism, the algorithm will continue to present them with similar material, creating an echo chamber effect. This insular environment makes it harder for users to encounter perspectives that challenge their racial biases. A prominent example of this is YouTube's recommendation algorithm, which has faced criticism for steering users toward far-right and racist content due to its reliance on engagement-driven prioritization (Russell, 2023).

Finally, feedback loop bias occurs when biased content is continuously engaged with, causing algorithms to amplify it. For instance, if white users are more likely to interact with "race-neutral" content, the algorithm will continue to prioritize these narratives, thereby reducing the visibility of discussions on systemic racism. This feedback loop strengthens the dominant racial narrative while marginalizing voices that challenge or critique racial inequality.

Once the different types of algorithmic biases have been analyzed according to the categorization of both Mehrabi *et al*. (2021) and Silva and Kenney (2018), we proceed to highlight specific research on how these have reproduced and intensified racist stereotypes. One of the most prominent studies on the issue is that of Noble, who categorizes these tools as "algorithms of oppression," emphasizing their role in sustaining and even exacerbating biases that marginalized groups already experience in the physical world. She argues that this issue will become one of the most pressing human rights challenges in the future, as algorithmic decision-making strengthens oppressive social structures and introduces new forms of racial profiling. In her seminal work, Noble (2018, pp. 6-24) provides striking examples of Google's search engine displaying racist and misogynistic results. For instance, its image recognition system automatically labeled Black individuals as "apes" or "animals" and associated Michelle Obama with similar terms. Additionally, Google searches for the derogatory term “Nigger House” led users to the White House, while queries for “beautiful” primarily returned images of white women. Similarly, autocomplete suggestions for “women cannot” and “women should” reinforced gender stereotypes, including “women cannot drive”, “women should not vote”, and even “women should be slaves”. Google and other tech giants dismiss these issues, claiming they are not responsible for their algorithms’ decisions and that such incidents are resolved quickly (Noble, 2018, p. 159). Google, with its search engine, Chrome browser, YouTube, and Gmail, is often perceived as synonymous with the internet. This perception normalizes biased results, as users assume they are objective and reflective of popular opinion (Noble, 2018, p. 34). These platforms, due to their dominance, shape online reality.

Ultimately, Noble contends that search results reflect corporate interests and advertiser priorities, rather than an unbiased truth. Harmful stereotypes persist because they are widely circulated, normalized, and profitable, highlighting the urgent need for accountability in AI-driven platforms. These issues, far from being addressed, may be perpetuated by the recent policies of major companies behind social media platforms to reduce content moderation in an effort to eliminate racist discourse. X, with the arrival of Elon Musk, as well as Meta, Microsoft, and Google, have downsized their teams dedicated to consumer and user well-being, including those focused on racial equity (Noble *et al*., 2025, p. 271).

4.3. The Role of Social Media in Spreading and Normalizing Racist Narratives

Social networks have become omnipresent channels of communication and dissemination of all kinds of content. Their users have ceased to be passive recipients of information, as was traditionally the case with conventional media, to become true protagonists—or so they believe, since, as we have seen, algorithms limit free will and the existence of bots reduces their prominence—by having the ability to create and share content consumed and noticed by hundreds, thousands, or even millions of people across any given territory.

The traits inherent to social media—mainly, anonymity, the potential for a massive audience, and the immediacy of interactions—among other factors, create an environment particularly conducive to the spread and normalization of racist narratives. Anonymity allows users to express extreme views without fear of direct repercussions, emboldening individuals to articulate racist rhetoric they might otherwise suppress in offline spaces. Although, strictly speaking, users may not be entirely anonymous, as they can be traced, what matters is the *sense* of anonymity, which encourages them to shed the social filters necessary to comply with norms of civility and decorum.

Furthermore, the promise of reaching vast audiences with a single post incentivizes provocative content, while immediacy fosters impulsive reactions, often without time for reflection or critical engagement. There are no filters, unlike traditional media and their editorial controls. These features intersect with the logic of virality, a cornerstone of the attention economy, where the most controversial and emotionally charged content is algorithmically amplified. As a result, racism—particularly in its culturalist form—flourishes, repackaged as “political incorrectness” or “anti-woke” sentiment to skirt overt accusations of bigotry.

Algorithmically amplified discourse further entrenches racist narratives by privileging engagement over accuracy or ethics. Social media algorithms, designed to maximize user interaction, push divisive content to the forefront since outrage and controversy generate higher levels of participation. This dynamic not only increases the visibility of racist ideas but also creates a false sense of legitimacy, as frequently encountered content—regardless of its veracity—appears credible by virtue of repetition. Consequently, racism becomes normalized in the digital public sphere, subtly shifting societal boundaries of what is considered acceptable speech.

The interplay between echo chambers, extremism, and racial discourse compounds this effect. Social media algorithms curate personalized feeds based on user behavior, reinforcing pre-existing beliefs and isolating individuals within ideologically homogeneous spaces. These echo chambers facilitate radicalization by constantly exposing users to similar viewpoints, with little to no counterbalance from dissenting perspectives. Within such insulated environments, racist ideologies can evolve and harden, as users embolden one another and escalate their rhetoric in a competitive bid for recognition and approval.

Moreover, the blurred line between freedom of speech and hate speech in digital spaces complicates efforts to curb online racism. While social media platforms often claim to uphold free expression—rooted in the American understanding of freedom of speech as an almost absolute right—their reluctance to impose stringent content moderation policies allows racialized hate speech to masquerade as legitimate political opinion[5]. In fact, hate

[5] The United States Supreme Court, in its interpretation of the First Amendment of the Constitution, allows all types of speech, including racist content, as long as these two elements are not met together: that the speech is "directed to inciting or producing imminent lawless action" and the speech is "likely to incite or produce such action." *Brandenburg v. Ohio,* 395 US 444 (1969). This means that, unlike other regions such as Europe, racist speech has more freedom to persist, as it is only limited by those two onerous requirements and does not take into account other legal goods that require protection, such as dignity, a key element for authors like Waldron (2012). In contrast, the Inter-American and European human rights systems have adopted more interventionist approaches with regard to racist hate speech. The Inter-American Court of Human Rights, interpreting Article 13.5 of the American Convention on Human Rights, has accepted that states may restrict expressions that advocate racial hatred, provided that such limitations are established by law and meet the criteria of necessity and proportionality (e.g., *Norín Catrimán et al. v. Chile*, concerning speech that reinforced racial stereotypes against Mapuche leaders). The Special Rapporteur on Freedom of Expression of the Inter-American Commission stressed that racist discourse sustains structures of exclusion and discrimination, and called for comprehensive responses beyond criminal prosecution—including anti-racist public policies and oversight of digital platforms. The report notes that racist speech should be prohibited when it incites violence or discrimination, but not in cases of unequal or offensive opinions, where "the most appropriate response is the fairness of reasoned argument, which calls for more and better speech, not less" (CIDH, 2017, p. 157). Similarly, the European Court of Human Rights has ruled that racist hate speech is not protected under Article 10 of the European Convention and may be excluded entirely from protection under Article 17 when it contradicts the Convention's fundamental values (e.g., *Garaudy v. France*). The European Union, through instruments such as Council Framework Decision 2008/913/JHA, has obliged Member States to criminalize the public incitement to violence or hatred on the grounds of race, colour, descent, or national or ethnic origin, and more recently, the Digital Services Act imposes obligations on large platforms to act against the dissemination of illegal hate content. These frameworks reflect a shared view that the freedom of expression cannot be used to undermine the equal dignity of others, especially in racially plural societies.

speech has significantly increased since Elon Musk acquired Twitter and rebranded it as X (Hickey *et al*., 2024, p. 1135). This permissiveness emboldens culturalist racism, which strategically avoids biologically deterministic language in favor of coded references to cultural incompatibility, national identity, and "threats" to societal values. By eschewing overtly racist terminology, these narratives exploit the ambiguities of moderation policies, enabling racism to persist under the guise of open debate.

Finally, moderation policies themselves sometimes sanitize or whitewash racist discourse due to, among other reasons, the lack of staff, the language barriers faced by those employed to moderate, as well as the lack of understanding of the cultural context in which these narratives occur (Sorabji, 2021, p. 2018; Gongane *et al*., 2023, p. 129). Rather than dismantling the underlying ideologies, platforms may simply remove explicit slurs while allowing more subtle, culturally racist narratives to thrive. In doing so, they create the illusion of a "cleaner" online space without addressing the systemic spread of racialized misinformation and bigotry. This superficial approach ultimately reinforces the normalization of racism, allowing it to adapt and flourish within the digital public sphere.

With this approach, platforms have chosen to uphold an absolutist conception of freedom of speech, fostering and cultivating what Justices Holmes and Brandeis[6] characterized as a *marketplace of ideas*, a metaphor which philosophically has its roots in the defense of freedom of speech put forth by both Milton (1918) and Mill (1970): a space where all opinions compete and can be expressed and tolerated to allow truth to emerge. Consequently, restricting such freedom through legal means becomes challenging, leaving states as the only actors capable of limiting hate speech within their jurisdictions. However, the transnational nature of social media enables the dissemination of such content without barriers, even in jurisdictions where its expression is prohibited.

Nevertheless, contrary to the intentions of those thinkers and Justices Holmes and Brandeis—who envisioned that truth would prevail over falsehood through open confrontation—the reality is quite different. The nuances of truth do not fit within the spatial constraints imposed by certain platforms, and algorithms neither evaluate nor possess the capacity to assess truth. What they do recognize, however, is that truth, often less engaging than falsehood, fails to capture the attention of users eager for sensationalist content and of companies seeking to monetize that attention.

5. Conclusions

The public sphere has undergone drastic changes in recent years, with public opinion increasingly shaped by content disseminated in cyberspace and on social media platforms—essential vehicles for understanding this transformation and the ways in which the media agenda and public discourse are structured today. However, contrary to what the term public sphere might suggest, these are virtual spaces governed by companies that act as sovereign entities, leveraging this new domain for their own benefit. They do so by

[6] *Abrams v. United States*, 250 U.S. 616 (1919).

engaging users, making them believe they are free and that they personalize their experience through customization options (such as names, avatars, colors, and background images). In reality, however, users are guided by the invisible hand of algorithms, designed by these major corporations to monetize every moment of user engagement on the platform. The future of a deliberative public sphere and a healthy democracy will depend on regulatory interventions that promote transparency, accountability, and inclusive participation. As Pasquale (2016, p. 506) indicates, regulation "would improve the public sphere, not diminish it". Without proper oversight, these platforms risk becoming tools for ideological manipulation rather than spaces for open discourse.

How can users be kept on social media for longer periods, thereby increasing economic profit? By offering content that captures their attention and fosters addiction. What kind of content elicits this response? Emotionally charged, sensationalist, and misleading material. Where does racism fit into this equation? It is a form of emotional content, rooted in fear and ignorance, that exudes hatred and becomes particularly appealing to certain individuals. However, like the evolution of racism itself, this content can be subtle, allowing it to go unnoticed. Continuous exposure to similar content creates echo chambers that hinder calm and rational debate with those who hold different perspectives, reinforcing the perception of absolute truth among users who see only information that confirms their existing beliefs. Social media platforms ultimately contribute to the normalization of both covert and overt racist discourse, playing a crucial role in the resurgence of this form of discrimination against racialized others. They may do so deliberately—to generate controversy and increase engagement—or indirectly, without an explicit intent to cause harm. This is where algorithmic biases come into play: internal flaws within algorithmic processes that may perpetuate racism, either because they are trained on biased data or because they are designed by individuals who, consciously or unconsciously, incorporate their own biases.

Racism, now primarily manifesting in its culturalist form, is experiencing a resurgence in cyberspace, driven by corporate profit motives that shape user interactions within the new public sphere. This sphere appears neutral and open due to its accessibility and the absence of formal entry barriers. However, it conceals the private interests of these corporations, which, from the shadows, orchestrate a public discourse that normalizes and whitewashes racism. Ultimately, social media platforms, controlled by major corporations—many of which are based in the United States—advocate for an opportunistic defense of freedom of speech as an almost absolute right. This approach enables them to monetize the user experience by providing sensationalist, extreme, or even overtly racist content, fostering user engagement and prolonging their time on the platform.

To counteract these issues, users can combat algorithmic bias on social media by taking several proactive steps. Firstly, educating themselves about how algorithms work and recognizing potential biases is crucial. Diversifying information sources helps avoid echo chambers, while providing feedback to platforms can highlight and address biased content. Adjusting privacy settings to limit data sharing can reduce personal data exploitation, mitigating some biases. Supporting ethical tech initiatives and advocating for policies that promote fairness and transparency in artificial intelligence and algorithm design are also important. Participating in digital literacy programs enhances understanding of digital platforms and empowers users to make informed decisions. Advocating for greater transparency from tech companies about their algorithms can help hold them accountable[7]. Engaging in public discourse about algorithmic bias raises awareness and drives collective action. By joining conversations, writing articles, and participating in community discussions, users can highlight ethical challenges and contribute to broader movements for change. These actions collectively promote a more inclusive and equitable digital environment, ensuring that digital platforms serve as spaces for genuine deliberation and fair public discourse.

Policy-making should focus on mitigating algorithmic bias by promoting transparency in algorithmic processes and implementing measures to counteract biases in training data. Regulatory frameworks should mandate the disclosure of how algorithms operate and the criteria they use to prioritize content. Additionally, fostering diversity in the tech industry can help reduce the likelihood of biased outcomes. Diverse teams are more likely to identify and address biases that may be overlooked by homogeneous groups. Ensuring fair public discourse requires a concerted effort to reconcile the emancipatory potential of digital technologies with the need for ethical governance and institutional oversight. Recommendations include the development of ethical guidelines for algorithm design, the establishment of independent bodies to monitor algorithmic decision-making, and the promotion of digital literacy among users. Educating users about how algorithms work and the potential biases they may introduce can empower individuals to critically engage with digital content.

In conclusion, this study provides a comprehensive analysis of the interplay between social media, algorithmic bias, and racism, offering insights into the ethical and democratic challenges posed by digital technologies. By addressing these issues, work can be made towards a more inclusive and equitable digital environment that upholds the principles of democracy and social justice. The findings underscore the urgent need for regulatory interventions to mitigate algorithmic bias and ensure that digital platforms serve as spaces for genuine deliberation rather than tools for ideological manipulation.

This study contributes to the existing critical literature by explicitly linking algorithmic bias to the resurgence of cultural racism in digital environments, highlighting the economic motivations behind content amplification and the structural role of platforms in shaping discriminatory discourse. This work foregrounds the ethical implications of platform governance and its impact on democratic deliberation. It offers an interdisciplinary perspective that bridges media theory, algorithmic accountability, and critical race studies to propose concrete policy and civic responses.

[7] In line with this recommendation, articles 14 and 27 of the Digital Services Act require platforms to disclose in their terms and conditions the tools used for content moderation, including algorithmic decision-making, as well as to publish the main parameters of their recommender systems. These parameters must be made available to users in a way that allows them to modify the default settings.

Funding

This article has been produced as part of the R&D&I project PID2023-147896OB-I00, Multihuri, funded by the Ministerio de Ciencia, Innovación y Universidades.

Bibliography

AGUILAR IDÁÑEZ, M. J. and BURASCHI, D. (2016). "Del racismo y la construcción de fronteras morales a la resistencia y el cambio social: la sociedad civil frente a las migraciones forzosas", *Servicios Sociales y Política Social*, *111*, 29–44.

AGUILAR IDÁÑEZ, M. J. and BURASCHI, D. (2019). "Racismo «democrático» y fronteras morales: ¿Cómo construir una ciudadanía insurgente?". In SOLANES, Á. (Dir.), *Discriminación, racismo y relaciones interculturales* (pp. 155–188). Cizur Menor: Thomson Reuters Aranzadi.

AICHNER, T., GRÜNFELDER, M., MAURER, O. and JEGENI, D. (2021). "Twenty-Five Years of Social Media: A Review of Social Media Applications and Definitions from 1994 to 2019", *Cyberpsychology, Behavior and Social Networking*, *24*(4), 215–222.

ARENDT, H. (2006). *Sobre la violencia*. Madrid: Taurus.

ARIDOR, G., JIMENEZ-DURAN, R., LEVY, R. and SONG, L. (2023). "The Economics of Social Media", *Journal of Economic Literature*, *62*(4), 1422–1474. https://doi.org/10.1257/jel.20241743

BALIBAR, É. (2005). Violencias, identidades y civilidad para una cultura política global. Barcelona: Gedisa.

BANTON, M. (1996). "The Cultural Determinants of Xenophobia", *Anthropology Today*, *12*(2), 8–12.

BARBERÁ, P. (2020). "Social Media, Echo Chambers, and Political Polarization". In PERSILY, N. and TUCKER, J. A. (Eds.), *Social Media and Democracy. The State of the Field and Prospects for Reform*. New York: Cambridge University Press.

BAROCAS, S. and SELBST, A.D. (2016). "Big Data's Disparate Impact", *Calif. L. Rev.*, *104*, 671–732. http://dx.doi.org/10.15779/Z38BG31

BERNAYS, E. (1947). "The Engineering of Consent", *The Annals of the American Academy of Political and Social Sciente*, *250*(1), 113–120. https://doi.org/10.1177/000271624725000116

BHARGAVA, V. R. and VELASQUEZ, M. (2021). "Ethics of the Attention Economy: The Problem of Social Media Addiction", *Business Ethics Quarterly*, *31*(3), 321–359.

BIZUMIC, B. and DUCKITT, J. (2012). "What Is and Is Not Ethnocentrism? A Conceptual Analysis and Political Implications", *Political Psychology*, *33*(6), 887–909.

BLUM, L. (2002). "I'm Not a Racist, But...": The Moral Quandary of Race. Ithaca: Cornell University Press.

BRATT, C. (2022). “Is It Racism? The Belief in Cultural Superiority across Europe”, *European Societies*, *24*(2), 207–228.

BUCHER, T. (2018). *If... then: Algorithmic power and politics*. Oxford: Oxford University Press.

BURKE, P. (2023). *Ignorancia. Una historia global*. Madrid: Alianza Editorial.

CABRERA ALTIERI, D. H. (2021). “El algoritmo como imaginario social”, *Zer*, *26*(50), 125–145. https://doi.org/10.1387/zer.22206

CARR, C.T. and HAYES, R. (2015). “Social Media: Defining, Developing, and Divining”, *Atlantic Journal of Communication*, *23*(1), 46–65. https://doi.org/10.1080/15456870.2015.972282

CIDH (2017). Informe Anual de la Comisión Interamericana de Derechos Humanos 2016. Volumen II: Informe de la Relatoría Especial para la Libertad de Expresión. OEA/Ser.L/V/II, Doc. 22/17, 15 de marzo 2017.

CINELLI, M., MORALES, G. D. F., GALEAZZI, A., QUATTROCIOCCHI, W. and TARNINI, M. (2021). “The echo chamber effect on social media”, *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118.

DE LUCAS, J. (2003). Globalización e identidades. Claves políticas y jurídicas. Barcelona: Icaria.

FLORIDI, L. (2024). *Ética de la inteligencia artificial*. Barcelona: Herder.

GARCIA, M. (2016). “Racist in the Machine”, *World Policy Journal*, *33*(4), 111–117. https://doi.org/10.1215/07402775-3813015

GILLESPIE, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. New Haven: Yale University Press.

GONGANE, V. U., MUNOT, M. V. and ANUSE, A. D. (2023). “Detection and moderation of detrimental content on social media platforms: Current status and future directions”, *Social Network Analysis and Mining*, 12, Article No. 129. https://doi.org/10.1007/s13278-022-00951-3

GONZÁLVEZ, V. (2011). “Educación para la ciudadanía democrática en la cultura digital”, *Comunicar*, *36*, 131–138. https://doi.org/10.3916/C36-2011-03-04

HABERMAS, J. (1986). Historia y crítica de la opinión pública. La transformación estructural de la vida pública. México D.F.: Editorial Gustavo Gili.

HABERMAS, J. (1989). “The Public Sphere: An Encyclopedia Article”. In BRONNER, S. E. and KELLNER, D. M. (Eds.), *Critical Theory and Society. A Reader* (pp. 136–142). New York: Routledge.

HABERMAS, J. (2023). A New Structural Transformation of the Public Sphere and Deliberative Politics. Cambridge: Polity Press.

HANNON, M. (2023). “Public discourse and its problems”, *Politics, Philosophy & Economics*, *22*(3), 336–356. https://doi.org/10.1177/1470594X221100578

HERNÁNDEZ MORENO, N. (2022). "El mito de la identidad y sus supuestos enemigos: ficciones que marcan el trato diferenciado en la acogida de personas migrantes en Europa", *Trayectorias Humanas Transcontinentales*, *8*, 35–50. https://doi.org/10.25965/trahs.4701

HICKEY, D., SCHMITZ, M., FESSLER, D., SMALDINO, P. E., MURIC, G. and BURGHARDT, K. (2024). "Auditing Elon Musk's Impact on Hate Speech and Bots", *Proceedings of the International AAAI Conference on Web and Social Media*, *17*(1), 1133–1137. https://doi.org/10.48550/arXiv.2304.04129

JUST, N. and LATZER, M. (2016). "Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet", *Media, Culture & Society*, *39*(2), 238–258. https://doi.org/10.1177/0163443716643157

KIM, B. (2020). "Effects of Social Grooming on Incivility in COVID-19", *Cyberpsychology, Behavior and Social Networking*, *23*(8), 519–525. https://doi.org/10.1089/cyber.2020.0201

LEE, C., GLIGORIĆ, K., KALLURI, P. R., HARRINGTON, M., DURMUS, E., SANCHEZ, K. L., SAN, N., TSE, D., ZHAO, X., HAMEDANI, M. G., MARKUS, H. R., JURAFSKY, D. and EBERHARDT, J. L. (2024). "People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise", *PNAS 2024*, *121*(38), e2322764121. https://doi.org/10.1073/pnas.2322764121

LEE, N., RESNICK, P. and BARTON, G. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Institution.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K. and GALSTYAN, A. (2021). "A survey on bias and fairness in machine learning", *ACM computing surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.48550/arXiv.1908.09635

MENCZER, F. and HILLS, T. (2020). "Understanding how algorithms and manipulators exploit our cognitive vulnerabilities empowers us to fight back", *Scientific American*, December, 54–60.

METZLER, H. and GARCIA, D. (2023). "Social Drivers and Algorithmic Mechanisms on Digital Media", *Perspectives on Psychological Science*, 19(5), 735–748. https://doi.org/10.1177/17456916231185057

MILL, J. S. (1970). *Sobre la libertad*. Madrid: Alianza.

MILLER, A., ARNDT, S., ENGEL, L. and BOOT, N. (2021). "Nature conservation in a digitalized world: echo chambers and filter bubbles", *Ecology and Society*, *26*(3), 11. https://doi.org/10.5751/ES-12549-260311

MILTON, J. (1918). *Aeropagitica*. Cambridge: Cambridge University Press.

NARAYANAN, A. (2023). *Understanding Social Media Recommendation Algorithms*. Knight First Amendment Institute, Columbia University.

NOBLE, S. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

NOBLE, S., ROBERTS, S., BUI, M., BROCK, A. and SNOW, O. (2025). "Structural Racism in Tech: Social Media Platforms, Algorithmic Bias, and Racist Tech". In CHRISTAKIS, D. and HALE, L. (Eds.), *Handbook of Children and Screens. Digital Media, Development, and Well-Being from Birth Through Adolescence* (pp. 269–274). Springer.

OPHIR, E., NASS, C. and WAGNER, A. D. (2009). "Cognitive control in media multitaskers", *Proceedings of the National Academy of Sciences*, *106*(37), 15583–15587.

PASQUALE, F. (2016). "Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power", *Theoretical Inquiries in Law*, *17*(2), 487–513. http://dx.doi.org/10.1515/til-2016-0018

PERKOWITZ, S. (2021). "The Bias in the Machine: Facial Recognition Technology and Racial Disparities", *MIT Case Studies in Social and Ethical Responsibilities of Computing*, no. Winter 2021 (February). https://doi.org/10.21428/2c646de5.62272586

RODRÍGUEZ-IZQUIERDO SERRANO, M. (2017). "Hate Speech y sociedad de la información: la difusión del odio en Internet y las redes sociales". In ALONSO SANZ, L. and VÁZQUEZ ALONSO, V. J. (Dirs.), *Sobre la libertad de expresión y el discurso del odio. Textos críticos* (pp. 129–144). Sevilla: Athenaica Ediciones.

RUSSELL, A. (2023). *YouTube Video Recommendations Lead to More Extremist Content for Right-Leaning Users, Researchers Suggest*. Available at (last access: 21 March 2025): https://www.ucdavis.edu/curiosity/news/youtube-video-recommendations-lead-more-extremist-content-right-leaning-users-researchers

SAITO, N. T. (2021). "Why Xenophobia?" *Berkeley La Raza Law Journal*, *31*, 1–25.

SARTORI, G. (1993). *¿Qué es la democracia?* México D.F.: Editorial Patria.

SEELIGER, M. and SEVIGNANI, S. (2022). "A New Structural Transformation of the Public Sphere? An Introduction", *Theory, Culture & Society*, *39*(4), 3–16. https://doi.org/10.1177/02632764221109439

SILVA, S. and KENNEY, M. (2018). "Algorithms, Platforms, and Ethnic Bias: An Integrative Essay", *Phylon*, *55*(1–2), 9–37. https://doi.org/10.1145/3318157

SOLANES, Á. (2018). Derechos y Culturas: los retos de la diversidad en el espacio público y privado. València: Tirant lo Blanch.

SOLANES, Á. and HERNÁNDEZ, N. (2024). *Formas de combatir el racismo en las redes sociales*, València: Tirant lo Blanch.

SORABJI, R. (2021). "Free Speech on Social Media: How to Protect our Freedoms From Social Media that are Funded by Trade in our Personal Data", *Social Psychology & Policy*, *37*(2), 209–236. https://doi.org/10.1017/S0265052521000121

STOICA, A. A., HAN, J. X. and CHAINTREAU, A. (2020). "Seeding network influence in biased networks and the benefits of diversity". In *Proceedings of The Web Conference 2020* (pp. 2089–2098).

STOICA, A. A., RIEDERER, C. and CHAINTREAU, A. (2018). “Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity”. In *Proceedings of the 2018 World Wide Web Conference* (pp. 923–932).

SUNDSTROM, R. R. and KIM, D. H. (2014). “Xenophobia and Racism”, *Critical Philosophy of Race*, *2*(1), 20–45.

SUSSER, D., ROESSLER, B. and NISSENBAUM, H. (2019). “Online manipulation: Hidden influences in a digital world”, *Georgetown Law Technology Review*, *4*(1), 1–45.

WALDRON, J. (2012). *The Harm in Hate Speech*. Harvard University Press.

WANG, J., ZHANG, W. and YUAN, S. (2017). “Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting”, *Foundations and Trends® in Information Retrieval*, *11*(4–5), 297–435. http://dx.doi.org/10.1561/1500000049

WIEVIORKA, M. (1994). “Racismo y exclusión”, *Estudios sociológicos*, *12*(34), 37–47.

WISTRICH, R. S. (1999). *Demonizing the Other: Antisemitism, Racism and Xenophobia*. London and New York: Routledge.

Received: 3rd March 2025

Accepted: 8th July 2025