Amina Vatreš*
University of Sarajevo
Faculty of Political Sciences
Department for Communication Studies

THE AUTOMATION OF CONTROL: ALGORITHMIC CENSORSHIP AND THE CONSTRUCTION OF SUB-REALITIES

Original scientific paper
UDC 351.751.5:004.021
316.774:004.021
001.102-026.786.3
001.102-021.191
https://doi.org/10.18485/kkonline.2025.16.16.17

The paper argues that the explosion of communication channels and content volume has created an illusion of informed participation, while simultaneously enabling latent, multidimensional forms of censorship. Rather than approaching censorship as an explicit act of top-down act of suppression, mostly from centralized authorities, this paper explores what contemporary communication theorists define as *censorship through noise.* Closely tied to the phenomenon of information overload, this form of *smokescreening* emerges within the current digital information ecosystem, where abundance itself becomes a mechanism of suppression. Often going hand in hand with disinformation and so-called alternative facts, it further amplifies internal contradictions, entropy, and logical inconsistency; it blurs the line between verified reporting and fabricated narratives, producing an epistemic landscape where truth is drowned in a sea of falsehoods, and where multiple fragmented, decontextualized, and affectively charged "truths" proliferate. Ultimately, the paper introduces the concept of *dual-level censorship* - the coexistence of latent suppression through information saturation and explicit algorithmic content moderation. Through this lens, the study maps how the digital information ecosystem contributes to deepening social polarization, alienation, and a reconfiguration of the digital divide and the very concept of censorship itself.

Keywords: algorithms, censorship through noise, algorithmic censorship, filter bubble, filter clash, algorithmic gatekeeping.

1. Introduction

The idea of the Internet as a decentralized, democratic, and emancipatory public space (*cf.* Negroponte, 1995; Barlow, 1996) has long been regarded as a sphere resilient to censorship – especially under authoritarian regimes - transforming the public into a digitally networked public sphere - an intertwined online and offline space of interaction that continuously reconfigures itself and enables numerous networked social movements. As Manuel Castells (2000) argued, the foundation of the network society lies in technological and communicational interconnectedness. However, the contemporary experience of living in a networked, information-saturated age differs sharply from the early technoutopian visions that once surrounded the rise of the Internet. With the vast expansion of global digitalization and the overcrowding of online spaces due to the soaring number of users – primarily

* University of Sarajevo – Faculty of Political Sciences, Skenderija 72, 71000 Sarajevo, Bosnia and Herzegovina; e-mail: amina.vatres@fpn.unsa.ba

driven by the techno-capitalist economy of big-tech companies - the realities of today's digital ecosystems have fundamentally challenged these early optimistic assumptions. Instead of a global agora where diverse perspectives circulate freely, the idea of the digital sphere being an open and free public space has come into serious question, as many scholars argue that corporate platformization, surveillance capitalism, and algorithmic control have profoundly restructured the digital sphere (*cf.* Morozov, 2011; Zuboff, 2019; van Dijck, 2013; Noble, 2018; Gillespie, 2018). In other words, the algorithmically curated environments that prioritize personalization over plurality and the attention economy over civic deliberation have become markedly more evident and pervasive.

What emerges is a digital landscape not only created or shaped but also perceived as colonized by platform capitalism, with algorithms playing a central role – acting as gatekeepers of visibility and arbiters of relevance. We are witnessing the rise of fragmented audiences and clashing sub-realities, shaped and amplified by algorithmic operations within the digital sphere. Moreover, this shift signals not just a change in media consumption, but in a deeper epistemic transformation in how truth, knowledge, and credibility are produced, circulated, and contested - a process best understood through the lens of digital epistemology. Shaped by algorithms that govern content visibility - designed and trained within specific socio-political contexts and reflecting the political and economic interests of their creators – they are creating networks of meaning, influencing not only what people see but also what they consider credible or relevant, resulting in a fragmented epistemic landscape with significant implications for communication and democratic processes.

Moving beyond recent critiques of algorithmic mediation, this paper explores not only how algorithms function as technical filters but also as active agents of epistemic fragmentation and latent control. Specifically, it draws on the theoretical framework of *censorship through noise* (*cf.* Pomerantsev, 2019; Simon and Mahoney, 2022) seen as a form of content suppression achieved not through direct prohibition, but through saturation, distortion, and disorientation of information flows. Unlike explicit forms of censorship, the idea is that algorithmic gatekeepers operate silently, employing opaque, language-driven models to filter, decontextualize, or deprioritize information based on political, economic, or ideological parameters, ultimately blurring the boundaries between fact and fiction. Grounded in communication theory, sociology of technology, and media philosophy, the paper introduces the concept of *dual-level algorithmic censorship*: a specific synthesis of (1) latent censorship through algorithmic saturation and distraction, and (2) overt algorithmic censorship exercised via corporate or state-influenced control. An illustrative example of the first layer can be observed in moments of social and political upheaval, crisis, or imporant events, when users' feeds are being flooded with a vast volume of sensational yet unrelated posts –memes, celebrity scandals, viral challenges – strategically capturing attention, fishing and maximizing engagement, while drowning out critical political discussions.[1] The second layer is exemplified by the documented algorithmic suppression of

[1] In more institutionalized contexts, such as China, the strategy of flooding digital spaces with apolitical or irrelevant content – termed "reverse censorship" – has been documented as a deliberate state technique to suppress dissent through distraction. According to numerous analyses by scholars, journalists, and digital rights activists, this practice – also referred to as *astroturfing* – involves the covert dissemination of large volumes of fabricated or staged social media content, made to appear as the spontaneous expression of ordinary citizens (King, Pan, and Roberts, 2017).

pro-Palestinian content during the 2023 crisis**,** where content moderation practices were reportedly shaped by geopolitical pressures and corporate interests, thus illustrating the operational logic of digital control in practice. Importantly, these two layers of algorithmic censorship often operate *simultaneously* – with distraction and saturation providing the cover for more explicit, targeted acts of suppression – thereby reinforcing each other and further fragmenting the possibility for meaningful public deliberation.

Rather than positioning users solely as passive victims of algorithmic manipulation, this paper also interrogates the role of digital subjectivity and affective participation in reproducing informational silos. In doing so, it contributes to ongoing scholarly efforts to *redefine censorship, digital divides, and the algorithmic construction of truth* in the post-digital era. While extensive literature has explored classical forms of censorship as state-driven mechanisms of content suppression, and more recently the algorithmic governance of digital spaces, there remains a lack of conceptual frameworks that account for the simultaneous and layered operation of both latent and overt algorithmic censorship. Most existing studies either emphasize content moderation and deplatforming as explicit control mechanisms or examine *filter bubbles and information overload* in isolation. This paper addresses this gap by introducing the concept of *dual-level* algorithmic censorship, theorizing how seemingly unrelated phenomena - such as digital distraction and politically motivated filtering – function together to shape visibility, polarize publics, and restructure the digital public sphere. To address this conceptual void, the paper introduces the notion of *dual-level algorithmic censorship* as a framework for understanding how digital suppression operates both implicitly - through distraction, noise, and emotional saturation - and explicitly - via algorithmic filtering shaped by corporate and political agendas. Accordingly, the structure of the paper reflects this duality: the first analytical section (parts 3, 4, 5, and 6) focuses on the latent layer of censorship, exploring how information overload, affective clustering, digital tribalism, and algorithmic identity profiling contribute to epistemic fragmentation and reconfigure the *digital divide.* The second analytical section investigates the explicit layer of algorithmic control, including documented instances of content suppression, platform governance practices, and the entanglement of algorithmic moderation with geopolitical and ideological agendas. This two-part structure enables a holistic reading of how algorithmic infrastructures simultaneously shape perception, visibility, and knowledge production – constituting a new paradigm of censorship that is both subtle and systemic.

2. Methodological Approach

Aiming to critically examine the socio-political dynamics of digital censorship in the context of algorithmic governance, this paper employs a qualitative, theory-driven analytical approach grounded in critical media and communication studies. Such an approach is essential for unpacking the persistence and transformation of censorship in the digital age, where classical (explicit, state-based) censorship has not disappeared but now often operates simultaneously with more latent, algorithmically mediated forms**.** Rather than treating censorship merely as a legal or technological issue, this paper approaches it through a *communication studies standpoint*, emphasizing the importance of mediated meaning-making, symbolic power, and visibility politics**.** The communicological dimension is crucial because censorship does not simply block access to information; it actively transforms communicative

environments, redefines what becomes sayable and visible, and affects how identities, solidarities, and publics are constructed and contested**.**

To enable such an analysis, the paper draws from a multidisciplinary foundation - including sociology of technology, critical data studies, media philosophy, and political theory – but remains rooted in the normative and analytical concerns of communication research. This allows for a *deconstruction of technodeterminist narratives* that portray algorithmic systems as neutral or autonomous, and instead foregrounds the human, political, and economic interests embedded in their design and operation. Based on this framework, the paper introduces the concept of *dual-level algorithmic censorship**,*** grounded in existing theoretical work on *censorship through noise*, but extends it by theorizing how latent distraction and overt suppression function simultaneously within algorithmically structured public spheres. This original contribution offers a *communication-centered mode*l for understanding how platform logics shape public visibility, epistemic fragmentation, and the reproduction of power in digital spaces.

Using a form of theoretical sampling by selecting key scholarly concepts, discourses, the analysis adopts an interpretive analysis of scholarly literature, policy reports, and platform-specific case studies, with particular attention to how algorithmic architectures shape information flows and public discourse., synthesizing diverse insights into a coherent framework for understanding contemporary informational ecosystem. This qualitative design relies on discourse analysis and the strategic selection of illustrative examples based on their analytical relevance to the socio-communicative implications of algorithmic personalization, digital subjectivity, and emerging forms of censorship.

The aim of this sampling is not statistical generalization but rather conceptual illumination of the communicative mechanisms through which algorithmic infrastructures shape visibility, structure public discourse, and reproduce power asymmetries in digital environments**.** By drawing on well-documented cases of automated content suppression driven by biased agendas such as the weaponisation of visibility by silencing of pro-Palestinian content on social media platforms, this approach allows for the flexibility to identify broader patterns and trends in how digital infrastructures mediate visibility, regulate dissent, and reshape the boundaries of public discourse

3. The Automatization of Censorship

Conventional censorship traditionally refers to institutional forms of control over content - most visibly exercised by the state to suppress politically undesirable information and regulate what individuals or society at large can access, but also manifested through religious, moral, and corporate mechanisms that shape and restrict what can be publicly expressed or known (Jansen 1988). This form of censorship using bans, restrictions, and surveillance, extends beyond the media, also including visual arts, literature, and other domains of intellectual and creative expression, often shaping the boundaries of public discourse by silencing dissent and marginalizing alternative narratives. Historically, the media have been particularly significant targets of such control due to their role in shaping public opinion and mediating access to political information and social critique. As the central channels through which ideas circulate and narratives are legitimized, they have often been subjected to regulatory pressure, concentrated ownership, and political interference – tools frequently used to suppress dissenting voices

and sideline oppositional perspectives. Governments and corporate actors alike have recognized the power of media not just to disseminate information, but to construct reality itself – making them a strategic site for controlling dissent and marginalizing oppositional voices (Curran and Seaton, 2009; Herman and Chomsky, 1988). As Wendy Brown (2015) argues, the disciplining of public discourse often involves not just silencing, but reconfiguring the terms of what can be said, who can say it, and under what conditions. It is therefore unsurprising that, with the advent of the internet, it was initially perceived as a space with emancipatory potential – as an unregulated and free space facilitating faster flows of information, broader access to diverse sources, and expanded opportunities for civic mobilization, operating beyond the reach of the traditional gatekeeper of state legislation. As Castells (2012) emphasizes, digital networks played a critical role in enabling social movements in authoritarian contexts, while Tufekci (2017) notes that platforms like Twitter and Facebook were crucial in amplifying dissent and coordinating protest during uprisings such as the Arab Spring.

However, as authoritarian governments have increasingly responded to this potential by imposing digital restrictions, such as temporarily cutting off internet access or blocking specific social media platforms, it blurred the line between internet and freedom. In Turkey, Twitter – once an essential tool for communication and grassroots mobilization – was repeatedly shut down by the Erdoğan government during politically sensitive moments. Notably, ahead of the 2023 presidential elections, access to the platform was restricted, echoing earlier examples such as in 2014, when Twitter was blocked following the Taksim Square protests and just before local elections in Istanbul. The crackdown coincided with growing public dissent over a controversial internet law and the circulation of corruption allegations implicating members of Erdoğan's inner circle. Erdoğan, on the other side, sees the eradication of Twitter as a move in which "…everyone will witness the power of the Turkish Republic" (Dockterman, 2014). Still, viewing the relationship between censorship and the internet solely through the lens of state-imposed shutdowns is reductionist. First, authoritarian regimes like China, Russia, or Israel – seen by the German authors Ingo Dachwitz and Sven Hilbig as the global leader of surveillance industry (2025, 186) – strategically embrace technological advancements across both software and hardware domains to exert control over their populations, using digital tools to monitor behavior, track activities, and influence public opinion. This aligns with the concept of *digital authoritarianism* (Roberts and Oosterom 2024), which describes "…a way for governments to control their citizens (…) inverting the concept of the internet as an engine of human liberation" (Shahbaz, 2018). Consequently, censorship today often operates through more sophisticated mechanisms such as user profiling, algorithmic filtering, and behavioral nudging, particularly within authoritarian contexts.

On the other hand, such mechanisms are not limited to authoritarian regimes where companies often develop these systems directly for state use – most notably in China through the so-called 'Great Firewall' sophisticated apparatus of internet control, censorship, and surveillance – but are also present in liberal democratic contexts. Therefore, although new digital media may seem harder to censor – unless legally banned – control is not necessarily exercised directly by the state but is instead embedded in the very back-end architecture of digital platforms pushing users into polarized echo chambers that not only amplify *digital divides* but also serve as highly profitable business models. Due to the dynamic

nature, speed, and vast volume of content being created, shared, and consumed on digital platforms, they also actually facilitate the rise of disinformation while enabling more latent and systemic forms of censorship that often remain invisible to end users. The internet, therefore, is not only capable of destabilizing authoritarian regimes but can also disrupt democratic societies – as evidenced by the Cambridge Analytica scandal or the revelations of Facebook's whistleblower Frances Haugen, exposing how platform capitalism fuels hostility and outrage for engagement and profit. Notably, certain major Big Tech companies – such as Microsoft, Amazon, SpaceX, or Palantir – have not only forged extensive contracts with the US government and military but have also embedded their executives within public institutions (cf. Bria 2025; Bremmer 2021). This revolving exchange of personnel and infrastructure further blurs the line between commercial interest and state power, transforming private platforms into instruments of governance.

4. Mass Audience from Networked to Fragmented Society

In moments of global crisis, such as the COVID-19 pandemic, a simple online search was often enough to confirm the widespread presence of shared concerns, public health responses, and political tensions across the globe. Yet, paradoxically, on users' individual social media feeds, the visibility of these events varied dramatically – not in terms of their occurrence, but in terms of *how* they were framed, *which* aspects were emphasized, and *which* details were either amplified or omitted. Depending on one's prior browsing history, emotional engagement, ideological leaning, or even geographic location, users encountered a fragmented set of narratives. The more interconnected we became, the more we were algorithmically steered into informational pockets aligned with our perceived preferences – shaped by our clicks, shares, comments, or viewing behavior, together with thousands of other users' patterns. In this way, the very affordances of digital platforms – interactivity, personalization, and algorithmic tailoring – turned into mechanisms for selective exposure. What emerged was not a universally accessible information space, but a system of segmented, affectively charged micro-realities. The pervasive interconnectedness facilitated by digital technologies has led to a reconfiguration of the public sphere into a digitally networked public (Tufekci, 2007, 6) simultaneously resulting in the possibility of more precisely targeting interests and selectively accessing content that users will respond to. Interactivity, personalization, and participation in content production become determining factors of digital discourse as users actively shape their online presence and, in connection with that, their digital identities. In this context, reflecting on the repercussions of growing segmentation, differentiation, and social stratification, Castells (2000) predicts the end of the mass audience and introduces a theoretical shift towards the concept of a segmented or fragmented audience. Instead of focusing on a general, heterogeneous audience marked by common general interests, the emphasis shifts to an audience classified according to specific needs, resulting in the adaptation of information to the relevance of its particular segment.

However, this also raises questions about the exposure to diverse information, the filtering of opposing perspectives, and the long-term effects of digital technologies on forming real social communities. Reflecting on how new advanced technologies, with a particular emphasis on artificial

intelligence technologies, support the process of audience segmentation and content personalization leads to a highly significant phenomenon identified in media influence theories as the information cocoon or filter bubble, whose impact seems to be insufficiently recognized. The term personalization refers to the process of tailoring digital content to users through AI technology and data analysis techniques, a process that is reshaping the way we consume and interact with information.[2]

Personalization becomes evident even in daily use of digital platforms and web search engine behavior, which aligns with previously identified user behavior patterns and is aligned with individual preferences, behavior hithe story, demographics, and other relevant characteristics.

Comparing the experiential field of social media use to a tunnel, Eli Pariser (2011) views the filter bubble as a product of the algorithmic classification of people and content personalization, consequently narrowing users' viewpoints. By connecting us exclusively with individuals who share the same or similar ideological and political positions and sorting content according to specific patterns that reflect our previous reactions, the underlying functioning of social networks traps us within the confines of informational cocoons. In other words, our digital footprints in the virtual space colonize us within the boundaries of a metaverse of personalized content, fundamentally transforming how ideas and information are received. Introducing the phrase "friendly world", Pariser emphasizes how online algorithms create a digital environment that seems pleasant and familiar to users while simultaneously and latently isolating them from diverse perspectives through a reductionist approach to reality.

Furthermore, a direct implication of this is the limitation of critical thinking and exposure to diverse perspectives. However, it is essential to emphasize that it would be entirely incorrect to ignore the fact that we directly and often unconsciously contribute to the process of filtering reality by making certain choices, such as whom we have as friends on social networks, on which posts we linger, and what we react positively to. It is imperative to consider the fact that social media users are not merely passive recipients or consumers; on the contrary, they realize their activity through the production of their content. This phenomenon is further amplified by users' inherent engagement as essential labour in the digital environment. Simply put, we become the creators of the walls of our own prison.

In this context, it is crucial to emphasize how explicit and evident the content offered within our filter bubble is while the complexity of what is implicitly omitted is simultaneously suppressed. Based on this thesis, Bucher (2018) concludes that algorithms control the boundaries of content visibility and contribute to the unfounded simplification of complex narratives. This undoubtedly opens up space for profound sociological and political implications. For instance, social activism or engagement on networks,

[2] It is imperative to underscore the predominance of commercialized, market-oriented apologetics of digital personalization. Commercially driven digital personalization entails an approach within the framework of digital marketing and e-commerce aimed at tailoring user experiences based on their personal preferences, behaviours, and characteristics to create individualized interactions between the company and the consumer. Conversely, scholars such as Pariser (2011) advocate for the necessity of a holistic approach to interpreting the personalization of digital content, asserting that it would be exceedingly reductionist to conceptualize this phenomenon solely within the confines of targeted advertising, such as offering products and services based on users' previous searches. Pariser substantiates his thesis with the example of Facebook, whose personalized news feeds are becoming the primary source of information for an increasing number of its users. He concludes that "...the algorithms that govern our ads are beginning to govern our lives." (Pariser 2011, 8).

often praised in the early 21st century (such as Occupy Wall Street or the Arab Spring), can easily fall into such reductionism, often without the actors' awareness.

Following this line of thought, Pariser emphasizes that in the virtual space, specific socially relevant issues disappear entirely, and entire political processes are also suppressed. He concludes, "The filter bubble often blocks things in our society that are important but complex or unpleasant. It makes them invisible. Moreover, it is not just issues that disappear. Increasingly, the entire political process is disappearing." (Pariser, 2011, 46)

5. Digital Identities and Information Bubbles – Interconnected or Isolated?

Professor of Media Studies at the University of Virginia, Siva Vaidhyanathan, argues that Facebook and other social networks narrow our field of vision without expanding our knowledge but instead keeping it within the same political-ideological framework. According to him, "…discussions among users whose opinions differ often boil down to arguments about what is true and what is false. Within the Facebook platform, the conversation usually remains at the level of contradictory assertions."[3] An indicative example of this is the situation from a few years ago when former Facebook employee and whistleblower Frances Haugen revealed her findings about Facebook's operations and confirmed that the platform previously relied on editorial policies primarily aimed at polarization and exaggerating distinctions among users rather than on constructive dialogue and substantive connection among its vast user base.

Emphasizing that behind such policies predominantly lie profit motives rather than the promotion of the public good, Haugen presented internal documentation to support her claims illustrating these aspects. In this regard, this phenomenon can be understood as a state of intellectual isolation that arises as an implication of filtering and reducing reality. Characterizing social networks as actually anti-social, Vaidhyanathan problematizes the far-reaching impact of the filter bubble, raising the question of whether its consequence is merely limiting users' field of vision or whether it even "…tribally determines us even more than we do ourselves." (Vaidhyanathan, 2018, 98). This narrowing of perspective leads to what Jamie Bartlett (2018) calls "retribalization" – the fragmentation of stable identity into a series of partial affiliations driven more by affect than rational discourse. Users gravitate toward echo chambers and filter bubbles not to engage in critical exchange but to reaffirm pre-existing emotional affiliations. As a result, rather than deliberative solidarity, we get affective clustering around identity fragments. Bartlett (2018) warns that: "Tribalism is understandable, but ultimately harmful to democracy, as it exaggerates small differences among us and turns them into huge, insurmountable chasms" (*ibid.,* 49)".

5.1. Digital Alienation and the Collapse of Otherness

Interpreting the contemporary era as an era of post-democracy, Benasayag identifies the greatest challenge of the future as the state in which "…machines eliminate the Other" and poses the question

[3] The evidence included studies that unequivocally demonstrate that Facebook possessed data indicating that Instagram significantly contributes to mental health issues among teenage girls. Additionally, evidence suggested that Facebook intentionally misled investors through public statements that did not correspond with the company's internal actions (Paul and Milmo, 2021).

"How should we fight against the complete annihilation of otherness." Often unconsciously and mostly involuntarily, we are, figuratively speaking, trapped in algorithmic echo chambers of like-minded individuals, which inevitably results in an experience of social alienation. Turkle concludes that people are "…increasingly drawn to technologies that create the illusion of togetherness without the demands of real relationships."[4] (Turkle, 2011; according to McChesney, 2015, 14).

One of the most effective ways of suppressing opposing perspectives is through algorithms that promote unification and sameness. In his work *Filterworld* (2024), Kyle Chayka refers to this phenomenon as "pervasive flattening" – a form of algorithmic gatekeeping that reduces complexity in favor of uniformity. As the boundaries between the virtual and physical worlds blur, digital footprints multiply, further entrenching this homogenization.

Within this dynamic, there is a growing need to rethink the concept of the digital divide. Rather than viewing it only through the lens of access to infrastructure or technology, the divide today must be understood as shaped by deeper social and ideological antagonisms. Earlier understandings (*cf.* Lessig, 1999; Castells, 2000; Van Dijk, 2019) focused on ownership or access is no longer sufficient. Instead, the divide has shifted into a space where the ideological, cultural, and value-based contradictions of the analogue world are reproduced – and even intensified – online (Vatreš and Alispahić, 2024).

In other words, the digital divide today is not merely about who is online and who is not, but about which ideological camps individuals belong to, what narratives they are surrounded by, and how algorithms amplify these divisions. Algorithmic information cocoons don't just reflect societal polarization – they actively contribute to it. This reconceptualized divide is defined less by technological access and more by participation in fragmented, emotionally charged, and ideologically aligned digital publics.

5.2. Fragmented Selves and Algorithmic Profiling

If we understand subjectivity and identity as socially constructed categories, then interaction on social networks becomes a key space where layered identities are shaped, negotiated, and performed. In this context, Bosnian-Herzegovinian political scientist Esad Zgodić (2009) argues that what he terms the phenomenon of *anonymous publicness* implies the existence of plural, multidimensional, and segmented identities – identities that individuals simultaneously express through the anonymity offered by virtual space. Drawing on Mark Poster's concept of cyberdemocracy (1996), as cited by Zgodić (2009), the Internet enables postmodern constructions of fluid identity: invented, layered, and often fragmented. "The Internet enables postmodern constructions of identity. Here, identities can be invented and changed; elaborate descriptions of the self are played out." (Poster, 1996; according to Zgodić, 2009, 478)[5]

[4] For further discussion on this phenomenon, see: Turkle (2011), *Alone Together: Why We Expect More from Technology and Less from Each Other.*

[5] The thesis regarding the existence and potential manifestation of plural, segmented, and fluid identities within the digital sphere, particularly in the contextual framework of the aforementioned phenomenon of anonymity, can also be recognized in a recent announcement by Meta published at the end of September 2023. Meta announced that Facebook users could create multiple profiles, each reflecting exclusively selected aspects of their personality. This should be interpreted as a paradigm shift in understanding digital identities, implicitly acknowledging their characteristics such as multilayeredness, fluidity, and the possibility of self-determination. This even officially encourages the creation of fake user accounts (see more in Peters, 2023).

As mentioned above, the personalization of digital content presupposes identifying some form of user identity, upon which the process of individual filtering and adaptation of the vast array of internet-accessible information occurs. In a deep interview from 2019, published in the book *Tyranny of Algorithms: Freedom, Democracy, and the Challenges of Artificial Intelligence*, Miguel Benasayag emphasizes how "…algorithms operate based on micro-information collected *en masse* in the digital space, which, when connected and correlated, determine profiles." (Benasayag, 2019, 80) Identifying individuals' micro-behaviors in the virtual environment as immediate reactions and merely "…a set of fragmented digital footprints", Benasayag presents the standpoint that there are no solid connections between algorithmically constructed identities and the real motives and essential interpretations of user searches. People perform different selves depending on the context of interaction. He further notes that algorithms do not construct coherent identities, but rather profiles based on fragmented micro-behaviors – what we click, like, or linger on. These fragments are then recombined into predictive models that lack depth, nuance, or interpretive clarity. The result is not a holistic identity but a mirror of immediate digital reflexes.

5.3. Between Google and Facebook: Dual Logic of Identity Construction

Analyzing the process of digital personalization on the web, Eli Pariser (2011) emphasizes the divergent approaches of different stakeholders in identifying user identity. Different platforms operationalize identity in different ways. Google primarily relies on individual search histories, creating a sense of privacy and introspection. Users feel anonymous, free to explore topics they might avoid publicly.[6] Facebook's filtering system, on the other hand, is outward-facing; it prioritizes what we choose to share and how we interact publicly. Thus, Facebook reflects a *performed self,* shaped by visibility and affirmation, while Google captures a *concealed self,* formed in private, often unconscious interactions. These opposing logics create a fractured sense of who we are, as digital selves become increasingly disconnected from the layered realities of offline identity.

Ultimately, while Facebook curates a performed self shaped by public interactions, and Google enables a concealed self that explores identity facets under a veil of privacy, both logics complicate the relationship between algorithmic profiles and real-world subjectivity. Within this predominantly profit-driven digital discourse, it is important to emphasize that commercial identities often aim to mirror – or even fully absorb – the complex and layered realities of actual user identities. The algorithmic gaze does not see people – it sees data points. And in this gaze, identity becomes not a question of self-expression, but of predictive value within an ecosystem built for extraction, not connection.

Reconceptualization of the digital divide through the lens of identity, rather than solely through material ownership of technology, emphasizes the significance of social aspects in shaping the general notion of the digital divide. This approach transcends purely technical or economic dimensions of the problem, placing emphasis on broader social contexts that contribute to the formation of digital

[6] This subjective sense of security encourages users to explore a variety of topics, reflected in their searches that, given this presumed privacy, might not be investigated in a less private environment (such as the space of social networks). It is essential to consider that the perception of anonymity can have profound implications on user behaviour, contributing to the dynamics of individual identity manifestation within the virtual framework.

segmentations. This way, the shift in understanding the digital divide allows for its broader and more holistic interpretation as a sociological category – one that goes beyond technical connectivity and primarily addresses questions of digital belonging and how advanced technologies deepen mutual misunderstanding.

Without intending to delve into a deeper analysis and interpretation, it's important to note the various forms of solidarity that emerge in response to the proliferation of various forms of fake news, alternative facts, and similar phenomena that favor the development of so-called conspiracy theories. Supporters of these theories, gathered in the digital environment, find affirmation for their views and beliefs, further stimulating their activity. In this way, the digital divide, understood in the sense explained above, is formed in the interplay between mainstream science and conspiracy theories.

Dialectically, being super-connected (*cf.* Chayko, 2018), we are somewhat excluded from real and empirical connections, especially those contrary to previously formed viewpoints and beliefs deviating from alignment with our ideological position. The blind immersion in our echo chambers prevents us from considering what could potentially influence a change in our perception of reality, as well as confronting arguments that challenge our views and beliefs: "The choice of content offered to you narrows over time because the posts of friends and sites are usually politically consistent. As reading the news chosen by our friends becomes increasingly the way we learn about what is happening in the world and its problems, the prospects of finding information outside our group diminish, leaving us deaf and blind to counterarguments and different claims." (Vajdijanatan, 2008, 20)

6. From Filter Bubbles to Filter Clashes – The Principle of Integrative Confrontation

Selective and narrow perception, bolstered by algorithmic filtering and reinforcing informational bubbles, significantly contributes to creating perceptual micro-realities aligned with various value beliefs and cognitive biases. However, some of the most recent scholars, such as Bernard Poerksen in his book *Digital Fever* (2022), have taken a bold step in reconceptualizing this phenomenon. They have marked the concept of a one-dimensional understanding of filter bubbles and the assumption of dystopian algorithmic isolation as outdated. Instead, Poerksen introduces a fresh perspective, a term he nominates as a *principle of integrative confrontation*, which challenges the traditional understanding of filter bubbles.

While not denying the existence of algorithmically conditioned fragmented micro-realities, he argues that even a single click represents a ticket to immersion in a completely different world, dissonant with our value orientations and ideologies, and consequently, a departure from the informational cocoon. However, the issue lies in the fact that despite the possibility of self-initiated reorganization of information flow and potentially consistent exposure to disparate perspectives, there still exist "…enclaves of information and perception, fragmented bubbles of the world and reality which may be determined algorithmically or primarily by social, cultural, or ideologically conditioned worldviews" (Poerksen, 2022, 96). The persistence of these phenomena is attributed to the fact that, although aware of the existence of our micro-reality established in relation to the other, we actively approach the other micro-reality by breaking through our filter bubble, not with the intention of critical reasoning and

questioning our positions, but with the receptive absorption of information that will support further confrontation.

In other words, although each new click potentially represents a doorway to other perceptual dimensions, netizens often open these doors exclusively with the intention of receiving information for the purpose of intentional conflict with the opposing views of another parallel world of (dis)information under the conditions of "homogenized diversity" (Bawden and Robinson, 2009). Just as in the *offline world,* exposure to pluralistic and opposing perspectives does not necessarily entail a critical reassessment of one's own cognitive biases. It has been demonstrated that such exposure in the virtual realm goes even further, contributing to greater polarization, as noted by Van der Linden (2024).[7]

In this context, the essential distinction from the analogue sphere lies in the digital transparency of diversity, which makes heightened tensions and direct or indirect confrontations between dissenters inevitable. Moreover, these confrontations are not limited to social status but extend to political-ideological worldviews that oppose each other in real time, conditioned by the context and fundamental dimensions of networked communication. It is noted that the "…networked world favours a mode of thinking akin to fragile fundamentalism", resulting in filter clashes as the inevitable, often explicit, conflict between "…parallel public spheres and self-affirming milieus" (Poerksen, 2022, 103). Furthermore, this environment generates a *new digital divide* understood as a sociological category that transcends technical connectivity, primarily centering on issues of digital belonging and illustrating the ways in which advanced digital AI technologies intensify mutual misunderstandings, as stated in Vatreš and Alispahić (2024).

7. From Conventional Censorship to Censorship through Noise

Paradoxically and seemingly counterintuitively, the formation and consolidation of digital informational cocoons occur under conditions of digital sphere overload, marked by *information overload*[8] and the constant active and passive reception of a multitude of information. This concept in the virtual domain entails an overload of information, misinformation, and alternative facts – defined as "…facts framed within a certain context or presented only partially to mislead the public and provoke a specific reaction" (Turčilo and Buljubašić, 2018, 7) – and often conspiracy theories, supported by logical inconsistency, entropy, and conflict between mutually contradictory and exclusive pieces of information. This reflects the most dangerous dimension of blurring the distinctions between fact-based reporting and entirely fabricated information, which, according to Simon and Mahoney (2022), consequently leads

[7] The outcomes of an experiment involving 1,600 Twitter users from both Democratic and Republican affiliations, who were exposed to divergent ideological viewpoints, demonstrated that augmenting the visibility of opposing perspectives is not always advantageous. „At the end of experiment, Republicans became more conservative in their attitudes (rather than more liberal) and Democrats became more liberal (rather than conservative). The end result was greater not less polarization." (Van der Linden, 2024, 127)

[8] Despite the absence of a singular, unified definition, we interpret information overload as a state in which an abundance of relevant, potentially useful, or entirely unusable information becomes more of an obstacle than an advantage to the recipient, hindering rather than enhancing the quality of informedness. Although this phrase does not exclusively refer to the digital realm, within the contextual framework of the online ecosystem, its meaning and crucial consequences become markedly more complex and, thus, a focal point of broader research interest.

to a kind of suffocation of truth in a sea of lies and the emergence of numerous fragmented, decontextualized, and subjective truths.

The virtual space of the internet, often framed within the discourse of technological utopianism as an entirely free and uncensored public domain characterized by a diversity of information and perspectives, is now facing a near-paradigmatic shift where the conventional meaning of censorship undergoes a genesis. To better understand how contemporary censorship as a concept manifests, authors Joel Simon and Robert Mahoney (2022), in their book *Infodemic*, emphasize that both the understanding and manifestation of this concept have fundamentally transformed in the modern digital age. They point out that censorship must now be considered within the contextual parameters of the virtual agora, a domain not typically associated with it in theoretical discussions.

In this way, modern censorship no longer relies on restricting and state-controlling the flow of information. However, it implicitly arises from information saturation and the existence of information cocoons, echo chambers, and filter clashes caused by an information blizzard. In this context, we can discuss an evolution of censorship, not only in its form but – more significantly – in interpreting its meaning and the complexity of contemporary mechanisms through which it is achieved. This phenomenon is closely related to information overload, one of the crucial determinants of the recent information-communication ecosystem, and the general clogging of communication channels. The social perception of knowledge and truth is implicitly shaped, distorted, and obscured, and instead of being factually grounded, it becomes affectively driven, subjectivized, and fragmented.The proliferation of new media and the quantity of available information thus create an illusion of plurality of opinions, which is particularly evident in the domain of social media and, more broadly, within the online sphere, where the issue of censorship acquires entirely new dimensions.

Instead of the conventionally understood concept of censorship as a method of explicitly blocking or limiting certain content from one or more centres, we refer to what recent mass communication theorists, such as Peter Pomerantsev (2019), describe with the term *censorship through the noise*. The ideology of restricting information gradually evolves towards an ideology of information abundance, indicating a latent genesis of censorship. However, this transformation must be understood not only as a change in approach to information accessibility but also as a strategic shift towards the instrumentalization of ubiquitous availability and the real-time dissemination of information as new means of censorship. In this sense, we are witnessing a transition from the information restriction paradigm to the information overload paradigm. According to Pomerantsev, this represents a war on reality, fundamentally supported by inauthentic online behaviours such as trolling, bot farms, and astroturfing, which intentionally pollute the public sphere.[9]

[9] According to Jelena Kleut (2020), it is necessary to make a clear distinction between the terms "bot" and "troll" since their use as synonyms has become common despite a clear differentia specifica in their semantic definitions. Although both terms represent components of online discourse pollution and digital manipulation, a troll is understood to be a participant in digital communication who "...deliberately uses impolite, aggressive, or manipulative statements to cause or deepen conflict, usually for their amusement" (Hardaker 2013 as cited in Kleut, 2018, 147). In contrast, a bot refers to an artificially constructed algorithm or software that performs pre-programmed automated actions, mimicking social interactions, and "...some bots are programmed to present themselves as humans" (*ibid.*, 148).

Consequently, numerous authors argue that this is a manifestation of a modern form of censorship that essentially achieves the same ultimate goals as traditional censorship but through different mechanisms. This complexity is further compounded by the fact that amidst the abundance of information, partial information, fake news, echo chambers, and information cocoons, it becomes challenging to draw a clear distinction between fact and fiction. Meanwhile, due to the illusion of ubiquitous information availability, we are not fully aware of the fundamental change in the way information is controlled in the online environment. According to Miroschnichenko (2020), the epistemological crisis, as one of the crucial determinants contained within the concept of a post-truth society, contributes to both the isolation of users and the polarization of society based on the clash of disparate subjective truths.

It could thus be asserted that the fundamental characteristics distinguishing the contemporary form of censorship, or censorship through noise, from conventionally understood censorship are reflected in at least three relevant dimensions:

- The inundation or overloading of communication channels with information rather than explicitly blocking or restricting access to it;
- The inability to clearly differentiate between factually based and entirely fabricated information;
- Information overload, the creation of filter bubbles, and clashes are not just incidental factors in the digital information environment. They are vital contributors and inherent conditions without which modern content censorship would not be as prevalent or effective.

At this point, the paper approaches the main thesis concerning the interrelation and (in)direct interdependence of the phenomena of information overload, the creation of filter bubbles and clashes as implications of filtration and personalization of online content on one hand, and the contemporary concept of censorship through noise on the other. With the escalation of the conflict in the Middle East, particularly the ongoing genocide Israel is commiting in the Gaza Strip, it becomes evident that censorship is not exhausted in its contemporary form. Rather, we are witnessing a partial return to traditional modes of censorship, differing only in the mechanisms of execution—now shaped by algorithmic conditioning and the logic of the new communication paradigm.

8. Algorithmic Censorship

It is evident that the synthesis of algorithmic training on large datasets, the vast amount of available information, and numerous invisible factors significantly influence the field and boundaries of visibility in the virtual domain. It is also important to note that, besides the users' activities, algorithmic classification and content personalization, which underpin the functioning of social networks, play a crucial and even decisive role in information filtering. This algorithmic filtering confines users within narrow echo chambers and information cocoons, creating a feedback loop perpetuating exposure to similar, cognitively biased content. As stated by Cobbe: "Algorithmic censorship extends the already extensive surveillance of the internet by commercial entities motivated primarily by profit. Out of this surveillance, social platforms can more actively and pre-emptively determine which speech should be

permitted and which should be suppressed, often according to their own criteria determined according to commercial considerations and incentives" (2021, 744).

Recent examples clearly indicate that algorithmically conditioned content filtering increasingly operates as a mechanism of digital control - *for instance*, by systematically marginalizing and obscuring pro-Palestinian narratives through blocking and shadowbanning - thereby moving beyond mere content moderation toward the active suppression of visibility. Consequently, there is a regression to traditional methods of content censorship, albeit through the use of different, AI-based invisible algorithmic barriers that narrow and direct the recipients' field of view in the digital space. This significantly affects the creation of recipients' perceptions of existentially relevant social issues. In such a communication environment, instead of journalists and media professionals, we are more likely to talk about *algorithmic gatekeepers* as creators of an algorithmically driven, often decontextualized, simplified, and one-dimensional sub-reality.

Demystifying the thesis of the neutrality of algorithms and artificial intelligence in general, Kostić emphasizes that "…throughout the entire process of development, research, design, testing, and implementation, various political, economic, social, and technological interests and agents are refracted through artificial intelligence and algorithms" (2021, 4). In this sense, it seems entirely justified to question whether, and if so, within what limits, we can speak of the Internet as a public space and a techno-utopian ideal of a space for free expression.[10]

The evolution and expansion of social media, and consequently the availability of online content, have undoubtedly led to a paradigm shift in content analysis, resulting in a transition from manual to automated moderation. This automated moderation, initially perceived as a utopian mechanism to combat misinformation and disinformation, intentionally deceptive narratives, hate speech, and similar negatively connotated phenomena within real-time online environments, has evolved into a crucial factor in what we now understand as algorithmic censorship. Thus, recognising established patterns, identifying specific writing styles, and semantic analysing phrases and keywords have become integral components of "training" algorithms to identify misinformation, disinformation, and even so-called conspiracy theories.

Algorithmic censorship presupposes the automated moderation of content through automated processes and algorithms of online platforms aimed at controlling, filtering, and restricting the dissemination of information on the Web. This form of censorship involves computational methods of regulating information flow and, as such, "…goes beyond traditional forms of censorship by involving automated decision-making systems, thus affecting content visibility. Such algorithms are designed to identify, promote, and suppress certain content based on predefined criteria" (Mohydin, 2023, 2). In this context, it is necessary to emphasize that the implications of algorithmic censorship should not be attributed solely to the reflection of freedom of expression. Rather it manifests prominently in the

[10] Problematizing the material dimension behind the operation of algorithms and the automated process of classification and selection of content in the virtual domain, thereby implying that the thesis of their neutrality is entirely obsolete, Hibert emphasizes that "...algorithms govern the interpassive illusion of free communication by concentrating the power of techno-corporate giants into autonomous geopolitical monopolies, whose labour force consists of the continuous living capital of networked citizens (netizens)." For further details, it is advisable to consult Hibert (2021).

broader context of an opaque access to information, the inability to disseminate it, and the deepening of existing digital polarizations and one-dimensional interpretations of social reality. Furthermore, the control of digital space is conditioned by the algorithmic *collapse of context*[11], which is associated with the reduced and simplified presentation of complex social phenomena within the framework of the digital information ecosystem.

The algorithmically conditioned content filtration on social media platforms has mainly been reflected in the case of censoring pro-Palestinian content related to the war and tragic civilian casualties in Gaza, once again debunking the thesis of algorithmic neutrality. However, this is not merely a coincidence without a complex background context. As a key player in the global surveillance industry, Israel was able to attract $8.1 billion in investments into its domestic cybersecurity sector in 2021 alone – driven in large part by the notorious *Unit 8200* of the IDF (Dachwitz and Hilbig, 2025, 186).[12] According to Mohydin (2023), established in 2015, this unit is responsible for telephone and electronic surveillance, operating in "…online surveillance and collaboration with social media platforms for content censorship." It is emphasized that this unit operates without clear legal procedures or user appeal mechanisms, focusing predominantly on areas such as gathering electronic information on cybercrime, managing criminal cases involving computer crimes, and implementing measures to combat prohibited online activities. Although ostensibly tasked with protecting the digital space, the emphasis of the Israeli Cyber Unit has been on posts allegedly supporting terrorist organizations or promoting incitement to violence.

Despite the evident lack of transparency regarding the criteria it uses when reporting content, the Israeli Supreme Court has ruled that the actions of this unit are lawful, "…revealing that Facebook complies with approximately 90% of the requests from this unit for the removal of Palestinian posts and accounts, with only 1% related to content involving racism, privacy violations, or other stated offences" (Mohydin, 2023, 5). Mohydin underscores that this strategy has been consistently used, as the unit has been actively removing content on social media alleged to incite violence associated with Hamas: "Since October 7, the same Israeli Cyber Unit has been actively removing content on social media alleged to incite violence associated with Hamas. The prosecution has filed approximately 4,450 removal requests, mainly directed at Meta as the parent company of Facebook, Instagram, and WhatsApp, as well as towards TikTok and X (Twitter)" (*ibid*.).

A recent report by Human Rights Watch addresses the same issue, problematizing the systemic censorship of pro-Palestinian content on Instagram and Facebook, based on its research conducted from October to November 2023, based on individual reports of online content censorship.[13] According to the findings presented in the report, Human Rights Watch identified six crucial patterns of unjustified

[11] In his book *Digital Fever: Taming the Big Business of Disinformation* (2022), Poerksen employs the term "collapse of context" to denote the underlying structural conditions of the emerging communication paradigm, which facilitate the formation of so-called filter bubbles—or, in his terminology, "filter collisions."

[12] Moreover, Israel's role in the field of cyber-surveillance is further evident – as noted by Antony Loewenstein – in its use of the Gaza Strip and the West Bank as testing grounds for new surveillance technologies, which are then marketed globally (Loewenstein, 2023). According to *Amnesty International* (2023), at least since 2019, the Israeli military has deployed the so-called "Wolf Pack" system, a biometric database containing extensive personal information on Palestinians from the occupied territories.

[13] The Human Rights Watch report is available at https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and (accessed February 11, 2024).

censorship, including removal of posts, stories, and comments; suspension or permanent removal of user accounts; restrictions on the ability to interact with content (such as liking, commenting, sharing); limitations on the ability to follow or tag other accounts; restrictions on specific functions, such as Instagram/Facebook live broadcasts, monetization, and account recommendations to non-followers; and shadowbanning, significantly reducing the reach of posts or accounts without prior notification to users.

One glaring example of explicit algorithmic censorship of pro-Palestinian content leads to the slogan "From the River to the Sea, Palestine will be free," under which protests in support of the Palestinian people have been organized worldwide. In hundreds of cases documented and analyzed by Human Rights Watch, this slogan, as well as comments such as "*Free Palestine*", "*IStandwithPalestine*", or "*Ceasefire Now*", were almost automatically removed from Instagram and Facebook user accounts without contextualization, attributed to "spam policy" community guidelines, despite the fact that the expressions themselves cannot reasonably be categorized as unwanted content or as violating the guidelines regarding incitement to violence, hostility, or discrimination in any way. The report also explicitly states that Meta has not offered a specific explanation for why the content and/or context in which these slogans appear justified their removal.[14]

Meta spokesperson Andy Stone spoke out on October 15, 2023, through a user account on the X[15] platform, attributing the reduced reach of user-generated content to a systemic bug, stating that the "…error equally affected all accounts worldwide and was not related to the content topic itself."

However, in the realm of shortcomings of algorithmically conditioned content filtration and its reliance on semantic analysis, keywords, and previously established repetitive patterns devoid of contextualization, space has been opened for pro-Palestinian activists to deceive the algorithms or engage in a sort of fight using the same weapons by which they were censored. When publishing content on social media, users intentionally avoid keywords by substituting them with similar ones or even omitting certain characters. These mechanisms of censorship evasion are encompassed by the concept of a new digital vocabulary labelled as algospeak, a portmanteau formed by the words "algorithm" and "speak", which refers to a coded language, "…a strategic tool for evading censorship on social media that involves creating alternative terms to replace keywords that could trigger automatic content moderation filters" (Mohydin, 2023, 6)[16]. Paradoxically, as algorithms increasingly shape linguistic expression and modes of communication, they also give rise to a distinctive form of digital resistance that challenges their own mechanisms of control.

One of the most relevant examples of the effective use of algospeak on social media, especially on TikTok, is using the watermelon emoji as a symbolic gesture of support for Palestine. This symbol has been used both to promote freedom of speech and a critical approach to the pro-Israeli narrative, as well as to organize fundraising campaigns to aid the Palestinian people. In this sense, algospeak has

[14] Additionally, one of the most frequently documented and illustrative indicators of unjustified censorship is the removal of the Palestinian flag emoji, which, according to Meta's terms of service, was unjustly attributed to the possibility of "...insulting, targeting, shaming, or harassing others."
[15] The statement is available at https://twitter.com/andymstone/status/1713603196080328741 (accessed February 15, 2024).
[16] For more in depth information, consult Holtermann (2023).

positioned itself and asserted itself as a central tool in responding to globally perceived biases in algorithm functioning, with an emphasis on the analyzed example of suppressing Palestinian voices from online public discourse. In this way, digital activism, supported by the effective application of algospeak, has actually proven to be a critical factor in documenting the tragic experiences of Palestinians, reporting from the ground, providing an alternative empirical perspective of reality, mobilizing part of the international support, and, generally speaking, in critically reviewing the Israeli official narrative.

9. Concluding Remarks

This paper set out to critically examine how algorithmic infrastructures shape contemporary forms of visibility, digital subjectivity, and emergent practices of censorship. By tracing the shift from mass to networked and fragmented audiences, and from filter bubbles to filter clashes, the analysis illustrated how algorithmic gatekeeping contributes to epistemic fragmentation, ideological polarization, and the distortion of democratic discourse.

Following the deconstruction of the techno-utopian paradigm of the internet as a completely free public space, in the contours of the online information ecosystem, we can now speak of *dual-level censorship* as a kind of synthesis of latent, implicit censorship through noise and the burden of communication channels, as well as algorithmic censorship as conventional censorship achieved through contemporary content filtration mechanisms. This *dual-level censorship,* almost paradoxically, occurs precisely in the conditions of interconnection and collision of information overload and fragmentation.

While algorithmic content moderation, as attempted to be demonstrated through this work, may result in the creation of filter bubbles in which users are isolated within their information bubbles, affirming their existing beliefs while simultaneously limiting access to divergent perspectives or facts, it also digitally socializes users and their fluid and multiple digital identities. In line with the existing algorithmic gatekeepers, established information bubbles, and echo chambers, bias in algorithmic content categorization can fatally contribute to solidifying existing social polarizations and ultimately, conflict of dissonant univocal perspectives.

This opens up space for both state and non-state actors, as well as corporate entities, to instrumentalize algorithmic censorship to manipulate narratives, suppress certain perspectives at the expense of others, and consequently shape public opinion in a controlled direction. It is essential to remember that algorithms should not be viewed solely as isolated entities but must be approached from the perspective of reflecting the socio-political environment from which they originate.

Crucially, this challenges the lingering technoutopian narrative of the Internet as a neutral or emancipatory space. It argues that algorithmic infrastructures operate on two intertwined levels of control: the first, latent and implicit, enacted through information overload, fragmentation, and personalized filtering; the second, more explicit, involving direct algorithmic moderation and political or commercial influence. This paper ultimately proposes the concept of *dual-level censorship* as a lens through which to better understand contemporary algorithmic power. By capturing both the implicit, noise-driven suppression of content and its explicit, algorithmically enforced filtering, this framework reveals the layered nature of digital control. In doing so, the paper expands the understanding of the

digital divide, reframing it not merely as a matter of technological access but as an ideologically and epistemically structured phenomenon. Rather than a question of who has access to digital tools, the divide now concerns who participates in which digital realities – and how those realities are curated, fragmented, and contested through algorithmic mechanisms. Recognizing this allows us to move beyond narrow definitions of censorship and begin to address the deeper epistemic and political consequences of algorithmic governance in shaping public knowledge, solidarity, and participation.

As the ultimate implication, there is also the deepening of the existing *digital divide,* which should no longer be understood merely as unequal access to resources but rather as a re-conceptualization oriented towards the existence of antagonistic ideological-political positions generated in the vacuum between information overload and algorithmic decontextualization and one-sidedness.

The colonization of online space by the flagships of corporate or platform capitalism, such as Facebook, Amazon, Google, or X, despite various attempts at regulation, leads to the training of algorithms according to corporate interests. In this way, they become instruments of manipulation, deliberate shaping, and control of public opinion, embedded within the logic of algorithmic capitalism - a system in which digital infrastructures, driven by profit-oriented algorithms, prioritize engagement, visibility, and monetization over truth and public interest. Consequently, these mechanisms risk reflecting and amplifying the personal, political, ideological, or corporate biases of their owners and stakeholders. This can result not only in deep social polarization but also in the dissemination of false information and the strengthening of alternative facts and conspiracy theories, which, under the guise of human rights and freedom of expression, become not only a threat to political stability but also to the very concept of factuality, truth, and knowledge. As distrust in government, political elites, and institutions grows, especially within developed liberal democracies, alternative facts and the authorities legitimized by them gain popularity. In other words, through these forms of censorship and the shaping of digital public opinion based on decontextualization, affective reactions, and emotions rather than facts, the foundation of democracy is lost.

The way out, therefore, does not lie merely in promoting algorithmic or digital literacy, but in fostering a holistic understanding of digital space, algorithms, and technology within the broader context of material conditions and its correlation with social and political power. Only by interpreting technology as embedded analogously in existing fields of conflict - as a terrain of political struggle for equality, participation, and freedom - can we work toward shaping the digital space not as a tool driven by corporate dominance, but as a public good and a genuine space of freedom.

References:

Amensty International. (2023, May 2). Isreali authorities are using facial recognition technology to entrench apartheid. https://www.amnesty.org/en/latest/news/2023/05/israel-opt-israeli-authorities-are-using-facial-recognition-technology-to-entrench-apartheid/

Barlow, J. P. (1996, February 8). *A declaration of the independence of cyberspace*. Electronic Frontier Foundation. https://www.eff.org/cyberspace-independence

Bartlett, J. (2018). *The people vs tech: How the internet is killing democracy (and how we save it)*. Ebury Press.

Bawden, D., and Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science, 35*(2), 180-191.

Benasayag, M. (2021). *The tyranny of algorithms*. Europa Editions.

Bennett, W. L., and Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication & Society, 15*(5), 739-768.

Bremmer, I. (2021). The Technopolar Moment: How Digital Powers Will Reshape the Global Order. Foreign Affairs, 100(6), 112-128.

Bria, F. (2025). Takeover by Big Tech. Le Monde Diplomatique, 2511, 2-3

Brown, W. (2015). *Undoing the demos: Neoliberalism's stealth revolution*. Zone Books.

Castells, M. (2000). *Uspon umreženog društva. Svezak I. Informacijsko doba: Ekonomija, društvo i kultura*. Golden marketing.

Castells, M. (2012). *Networks of outrage and hope: Social movements in the internet age*. Polity Press.

Castells, M. (2018). *Mreže revolta i nade: Društveni pokreti u doba interneta*. JP Službeni glasnik.

Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology, 34*(4), 739–766. https://doi.org/10.1007/s13347-020-00429-0

Curran, J., and Seaton, J. (2009). *Power without responsibility: Press, broadcasting and the internet in Britain* (7th ed.). Routledge.

Čejko, M. (2018). *Superpovezani*. Clio.

Dachwitz, I., Hilbig, S. (2025). *Digitaler Kolonialismus. Wie Tech-Konzerne und Großmächte die Welt unter sich aufteilen*. C.H.Beck;

Dockterman, E. (2014, March 21). *Turkey bans Twitter*. TIME. https://time.com/32864/turkey-bans-twitter/

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Herman, E. S., and Chomsky, N. (1988). *Manufacturing consent: The political economy of the mass media*. Pantheon Books.

Hibert, M. (2021). Podaci, algoritmi i umjetna inteligencija: Pismenost za 21. stoljeće. Univerzitet u Sarajevu – Fakultet političkih nauka. https://fpn.unsa.ba/b/wp-content/uploads/2021/03/Podaci-algoritmi-i-umjetna-inteligencija.pdf

Holtermann, C. (2023, December 27). Images of watermelons signal support for Palestinians. *The New York Times*. https://www.nytimes.com/2023/12/27/style/watermelon-emoji-palestine.html

Human Rights Watch. (2023, December 21). Meta's broken promises: Systemic censorship of Palestine content on Instagram. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

Chayka, K. (2024). *Filterworld: How algorithms flattened culture*. Heligo Books.

King, G., Pan, J., and Roberts, M. E. (2017). *How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument*. *American Political Science Review*, 111(3), 484–501.

King, G., Pan, J., and Roberts, M. E. (2017). *How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument*. *American Political Science Review*, 111(3), 484–501. https://doi.org/10.1017/S0003055417000144

Kleut, J. (2020). *Ja (ni)sam bot: Komentari čitalaca kao žanr participacije u digitalnom prostoru*. Univerzitet u Novom Sadu – Filozofski fakultet.

Kostić, B. (2021). Veštačka inteligencija: Uticaj na slobodu izražavanja, medijske perspektive i regulatorni trendovi. OSCE.

Lessig, L. (1999). *Code and other laws of cyberspace*. Basic Books.

Jansen, S. (1988). *Censorship: The Knot That Binds Power and Knowledge.* Oxford University Press.

Loewenstein, A. (2023). *The Palestine Laboratory: How Israel Exports the Technology for Occupation Around the World.* Verso.

Mekčejsni, R. V. (2015). *Digitalna isključenost: Kako kapitalizam okreće internet protiv demokratije*. Fakultet za medije i komunikacije Univerzitet Singidunum.

Miroshnichenko, A. (2020). *Postjournalism and the death of newspapers: The media after Trump: Manufacturing anger and polarization*. Toronto, Canada.

Mohydin, R. (2023). Algorithmic censorship and Israel's war on Gaza. TRT World Research Center. https://researchcentre.trtworld.com/wp-content/uploads/2023/12/Algorithmic-Censorship.pdf

Morozov, E. (2011). *The net delusion: The dark side of internet freedom*. PublicAffairs

Negroponte, N. (1995). *Being digital*. Vintage Books.

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

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin.

Paul, K., and Milmo, D. (2021, October 3). Facebook putting profit before public good, says whistleblower Frances Haugen. *The Guardian*. https://www.theguardian.com/technology/2021/oct/03/former-facebook-employee-frances-haugen-identifies-herself-as-whistleblower

Peters, J. (2023, September 21). Facebook officially embraces fake profiles. *The Verge*. https://www.theverge.com/2023/9/21/23883843/facebook-fake-profiles-finsta-meta

Poerksen, B. (2022). *Digital fever: Taming the big business of disinformation*. Palgrave Macmillan.

Pomerantsev, P. (2019). *This is not propaganda: Adventures in the war against reality*. Faber & Faber.

Roberts, T., and Oosterom, M. (2024). Digital authoritarianism: a systematic literature review. *Information Technology for Development*, 1–25. https://doi.org/10.1080/02681102.2024.2425352

Shahbaz, A. (2018). *Freedom on the net 2018: The rise of digital authoritarianism*. Freedom House. https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism

Schlogl, L. (2022). *Digital activism and the global middle class*. Routledge.

Simon, J., and Mahoney, R. (2022). *The infodemic: How censorship and lies made the world sicker and less free*. Columbia Global Reports.

Stone, A. (2023, October 17). [Tweet]. Twitter. https://twitter.com/andymstone/status/1713603196080328741

Terkl, Š. (2011). *Sami zajedno: Zašto očekujemo više od tehnologije nego jedni od drugih?* Clio.

Tolentino, V. (2019). *Trick mirror: Reflections on self-delusion*. Penguin Random House.

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Turčilo, L., and Buljubašić, B. (2018). Alternativne činjenice i post-istina u BiH: Ko (stvarno) kreira agendu medijima? United States Agency for International Development.

Vajdijanatan, S. (2018). *Antidruštvene mreže*. Clio.

Van der Linden, S. (2024). *Foolproof: Why we fall for misinformation and how to build immunity*. 4th Estate.

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

Van Dijk, J. (2019). *The digital divide*. Polity Press.

Vatreš, A., and Alispahić, S. (2024). Reconceptualisation of Social Solidarity: Networked Segmentation as a New Digital Divide. *Društvene i humanističke studije*, *9*(1 (25)), 939-962.

Zgodić, E. (2009). *Multiverzum vlasti*. Fakultet političkih nauka.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for ahuman future at the new frontier of power*. PublicAffairs.

Primljeno: 26. 7. 2025.
Prihvaćeno: 13. 10. 2025.