DOI: https://doi.org/10.36719/2663-4619/118/189-202

Nigar Garajamirli
SOCAR
https://orcid.org/0009-0008-8195-4629
nigar.garajamirli@gmail.com

Platform Governance in the Age of Media Created by Artificial Intelligence, Deepfake Technologies and Fact-Checking Challenges

Abstract

This article examines the changing nature of platform governance in the age of deepfake technology and algorithmic manipulation. It analyzes how social media platforms are adapting their disinformation policies, the implications for fact-checking and journalism, Meta's termination of third-party fact-checking partnerships, and how new governance mechanisms such as X's (formerly Twitter's) Community Notes are shaping public discourse. The study also evaluates ethical and educational responses to algorithmic disinformation by integrating international frameworks such as UNESCO's AI teaching guides, MediaSmarts' Break the Fake program, and WITNESS's Prepare, Don't Panic initiative. The aim is to evaluate platform policies not only from a technical and educational perspective, but also from a democratic participation, algorithmic bias, and social responsibility perspective. This study assessed the multi-layered impacts of synthetic media and deepfake technologies, encompassing not only technical but also political, social and ethical dimensions. Changes in fact-checking policies of social media platforms, digital attacks on journalists and user-generated content governance practices pose serious risks to democratic participation. Meta's end of its fact-checking partnerships and practices like the Community Notes program on the X platform have brought new conversations to combat misinformation, but the impartiality and effectiveness of these mechanisms are still in question.

Keywords: *deepfake technologies, disinformation policy, fact-checking, platform moderation, community notes, media literacy, algorithmic bias*

Nigar Qaracəmirli
SOCAR
https://orcid.org/0009-0008-8195-4629
nigar.garajamirli@gmail.com

Süni intellektlə yaradılmış media dövründə platforma idarəçiliyi, Deepfake texnologiyaları və fakt yoxlama çətinlikləri

Xülasə

Bu məqalə dərin saxtalaşdırma texnologiyaları və alqoritmik manipulyasiya dövründə platforma idarəçiliyinin dəyişən mahiyyətini akademik baxımdan araşdırır. Məqalədə sosial media platformalarının dezinformasiya siyasətlərini necə tənzimlədiyi, faktların yoxlanılması və jurnalistikanın bu sahədə üzləşdiyi çağırışlar, Meta-nın üçüncü tərəf fakt-yoxlama tərəfdaşlıqlarını dayandırması və X platformasında "İcma Qeydləri" (Community Notes) kimi yeni idarəetmə mexanizmlərinin ictimai diskursa təsiri təhlil edilir. Həmçinin, UNESCO-nun süni intellekt üzrə tədris bələdçiləri, MediaSmarts-ın "Break the Fake" ("Saxtanı Aşkar Et") proqramı və WITNESS-in "Prepare, Don't Panic" ("Hazırlıq Et, Panikaya Qapılma") təşəbbüsü kimi beynəlxalq çərçivələr inteqrasiya olunmaqla alqoritmik dezinformasiyaya qarşı etik və təhsil yönümlü yanaşmalar qiymətləndirilir. Məqsəd yalnız texniki və təhsil aspektlərini yox, həm də demokratik iştirakın genişləndirilməsi, alqoritmik qərəzin aradan qaldırılması və sosial məsuliyyət prizmasından platforma siyasətlərini obyektiv qiymətləndirməkdir.

Tədqiqat süni media və dərin saxtalaşdırma texnologiyalarının texniki, siyasi, sosial və etik səpkidə çoxşaxəli təsirlərini öyrənir. Sosial media platformalarında fakt-yoxlama siyasətinin dəyişməsi, jurnalistlərə yönəlik onlayn hücumlar və istifadəçi istehsalı məzmunun idarəolunma təcrübələri demokratik iştirak prinsipləri üçün ciddi risklər meydana çıxarır. Meta-nın fakt-yoxlama tərəfdaşlıqlarını ləğv etməsi və X platformasında "İcma Qeydləri" proqramının tətbiqi dezinformasiyaya qarşı mübarizədə yeni yanaşmalar təqdim etsə də, bu mexanizmlərin bitərəfliyi və effektivliyi hələ də sual altındadır.

Açar sözlər: *dərin saxtalaşdırma texnologiyaları, dezinformasiya siyasəti, fakt-yoxlama, platforma moderasiyası, İcma Qeydləri, media savadlılığı, alqoritmik qərəzlilik*

Introduction

Today, the transformation of digital media environments is shaped not only by technological advances, but also by structural changes in the production, circulation and control of information. Especially with the development of artificial intelligence-supported content production tools, the border between reality and fiction is blurring; This directly affects individuals' access to information, verification and evaluation processes. Among these technologies, "deepfake", which is the most noticeable, digitally imitates a man's voice or face, and allows him to attribute actions that he did not say or do in reality. Initially used in the fields of entertainment and art, this technology has caused serious ethical and social problems in recent years in areas such as political manipulation, gender-based violence and attacks on journalists.

Social media platforms, in particular, have become the main channel through which such manipulations are produced and spread rapidly. The active role of users in content production has placed responsibility on platforms not only as technical providers, but also as actors directing the flow of information. However, the algorithmic systems and community moderation models used by platforms in content control are subject to serious criticism in terms of both impartiality and transparency. User-based control tools such as Community Notes offer the possibility of collective access to information; coordinated abuse of these systems can accelerate the spread of false information instead of stopping it.

In this context, platform governance is not just a technical issue limited to blocking inappropriate content; it is also a complex structure that concerns basic rights such as freedom of expression, right to access to information, ethical responsibility and digital equality. Especially in countries in the Global South, this problem is felt more clearly; factors such as infrastructure, linguistic diversity and limited media literacy make these regions more vulnerable to misinformation.

This study aims to analyze the effects of artificial intelligence-supported content on the information ecosystem, especially through deepfake applications. In this framework, the governance policies, truth control mechanisms, algorithmic biases, and ethical responsibilities applied by social media platforms to combat misinformation will be evaluated as multi-layered. In addition, media literacy-based solution proposals of international actors such as UNESCO, MediaSmarts and WITNESS will be integrated into the analysis dimension of the study and a comparative evaluation will be made. Ultimately, the study is not just about technical solutions; at the same time, it aims to reveal how effective social and ethical approaches can be in the digital information system. In the conclusion, it is important to raise awareness of how dangerous deepfake technology can be in social and political terms now and in the future. It will open new discussions about the control of uploaded videos. Information will be provided on what precautions can be taken in this regard.

Research

In today's digital age, platform governance has ceased to be a mere technological issue, and has become a political, social, and epistemological problem. When social media platforms were first established, they advocated not to interfere with content and to protect freedom of expression. However, over time, this position began to change as the false information spread by users reached the power to create mass perception.

Digital platforms are no longer just tools for sharing information, but also the main channels for shaping social perceptions. AI-powered algorithms indirectly decide what individuals will see and when, while determining content ranking by analyzing user behavior. This process means institutionalization of "perception management" through technology. Especially this direction, carried out through seemingly harmless categories such as news recommendations, search results, recommended videos, deeply affects society's perception of reality.

Platforms such as YouTube, TikTok, Facebook and Instagram aim to provide more interaction by highlighting content that users are interested in. However, this interaction is often not related to the quality of the information, but to the triggering of emotions. Especially the content that evokes strong emotions such as anger, fear, and admiration is more prominent, which causes an increase in digital polarization. This is because algorithms only ask "what can we love?" not to the question, "why do we stay on the screen longer?" is related to the search for an answer to the question.

Artificial intelligence systems trap individuals in a "bubble of information" while serving content based on user history. This situation allows individuals to encounter only content that supports their views; thus, contact with different thoughts and ideas is reduced. This phenomenon is explained by the concepts of "echo chamber" and "filter bubble" and weakens the ground of democratic discussion in the digital environment.

Social media algorithms also influence political preferences, consumption habits, and cultural orientations. In a study conducted in India, it was observed that during election periods, content about certain parties or candidates is filtered according to the user's political inclination. This situation eliminates the chance for users to make an informed choice and prevents choices from being made on the basis of fair information.

On short video-based platforms such as TikTok, the content presented to the user is created in seconds; After watching only a few videos on a topic, the user encounters hundreds of similar content. This form of navigation has a high impact, especially on young users. Especially in sensitive topics such as mental health, sexuality, and body perception, the flow of content created by algorithms can turn into psychological pressure.

The impact of artificial intelligence systems on social perception is felt not only on the individual level, but also on collective preferences and public discourse. The presentation of deepfakes and similar synthetic contents on social media as if they were real causes the inclusion of artificial data in the formation of public opinion. This leads to a situation defined as "epistemic collapse", where doubt and uncertainty take the place of truth.

The way content is presented to the user on digital media platforms is no longer just for entertainment or access to information, but also for political and economic guidance. For example, in some countries, the algorithmic invisibility or late distribution of content calling for protest has become a tool that hinders social mobilization. This poses a serious threat to freedom of expression.

Most of the decision-making processes of artificial intelligence systems are closed to the user. Information is often not given about why the algorithm recommends one content and why it excludes another. This lack of transparency removes the ethical responsibility of the platforms and leads the user to a passive position.

Furthermore, since these systems are often trained with commercial datasets, they may carry cultural biases. For example, wrong generalizations are made based on data about gender roles or ethnic representation, which causes certain groups to be constantly represented in stereotypical ways. In this regard, artificial intelligence has the potential to reproduce social inequalities.

Otherwise, in the "algorithmic anarchy", individuals not only reach knowledge, but also submit to a guided reality.

The increase in violence against Rohingya Muslims, especially in Myanmar as a result of hate speech spread through Facebook, has shown the entire world that the passive attitude of platforms can lead to serious humanitarian consequences. After this incident, Facebook has reconsidered its understanding of governance and has turned to collaborations with algorithms and independent fact-checkers for content moderation.

Similarly, there have been lynching incidents in India as a result of fake news spread through WhatsApp, which has led to the emergence of new measures regarding the dissemination of content in messaging applications. Meta has imposed restrictions on the message transmission chain after these incidents.

These developments show that platform governance is not only a Western-centric problem, but a global one. For example, in Nigeria, Meta's withdrawal of third-party fact-checking partners led to the spread of disinformation during the local election period. Local NGOs have tried to fill this gap, but they have struggled to gain visibility without platform support (Gosse, Burkell, 2020).

In Indonesia, users organized by religious groups have provided mass negative feedback to content by women's rights advocates, ensuring that this content is suppressed by algorithms. Methods such as community rating systems or user tagging, while seemingly democratic in theory, can easily be instrumentalized in practice.

Platforms such as Wikipedia and Reddit have long used community-based content moderation systems, where volunteer users try to ensure accuracy. However, even these platforms are sometimes subject to coordinated attacks, making it difficult to remain neutral, especially on political issues.

Professional fact-checking organizations – such as Africa Check and Dubawa (West Africa) – operate under specific ethical rules and provide impartial information based on source analysis. However, since social media algorithms filter such content based on popularity, false information spreads much faster, while truth remains in the background.

Studies in the literature have shown that algorithmic moderation can be biased. For example, algorithms trained with sexist data may more frequently flag content from female users as 'inappropriate'. This situation does not only result in a violation of individual rights, but also in structural injustice.

In this context, the concept of "algorithmic justice" has gained importance. The decision-making processes of algorithms should be transparent, and it should be clearly stated to the user why content is removed or its visibility is reduced. The European Union's Digital Services Act (DSA) attempts to establish this principle of transparency in a legal framework.

However, technical reforms alone are not enough. Digital media literacy should increase users' capacity to evaluate content and create points of social resistance. As emphasized in UNESCO's "AI and Media Education" guide, media education should include not only tool knowledge, but also critical thinking, ethical evaluation and awareness of social responsibility.

Nevertheless, media literacy alone may be insufficient to combat information pollution. Especially in regions with low levels of education, perception management is much easier, and deepfake content spreads rapidly. The lack of actors who verify content in local languages in countries such as Niger and Chad has made these countries more vulnerable.

It is also common for people or institutions that do fact-checking to be discredited by political actors on the grounds of "bias". This strategy is used especially in authoritarian regimes to reduce the impact of fact-checking (Garg, Kumar, 2024).

On the other hand, technological developments also offer some solutions. The InVID project, developed in Europe, supports media professionals as verification software used in video analysis. In addition, open-source software helps civil society actors establish their own analysis systems by reducing costs.

The common point of all these studies is this: platform governance is multi-layered, and it is unrealistic to leave the solution only to platforms. States, civil society, technology developers and media actors should act together.

Today, with the developing technology, developments in video editing can allow the creation of videos that portray people doing things they have never done or saying words they have never said. Until recently, deepfake technology, which was only used by university researchers to create videos by experimenting, is now widely accessible with free or low-cost tools for producing fakes. In such an environment, situations may be created where corporate leaders and public relations experts will have difficulty convincing the public that a video is not real. As deepfake systems become more widespread, it can be thought that businesses may have difficulty convincing people that even real

videos are really real. In this context, the importance of recording videos of executives and other company spokespeople in public speeches will increase. In addition, it will be important to constantly and thoroughly examine social media posts that institutions are the subject of. It is necessary to determine whether the content is being sent or shared by false accounts. In this context, it is important to understand how social media algorithms work and how they can be directed against institutions (Cover, 2022).

In this context, various institutions have begun to develop projects that will control deepfake applications and help professionals detect fakes. For example, InVID, a research and innovation project supported by the European Union, has developed an information verification platform to detect emerging fake content and assess the reliability of news video files and content spread through social media. It is anticipated that investments in such projects will become increasingly widespread in the future and that these investments will become increasingly important (Rini, 2020).

Another important step will be for social media platforms to establish systems that will perform real-time fake detection in all their products and adapt to technological developments. In 2020, a fake voice recording produced using Deepfake technology was used as evidence in a custody case in the UK, and it has been practically proven how dangerous Deepfake technology is in negatively affecting judicial processes (Burkell, Gosse, 2019). If we look at the statistics, there has been a 350% increase in crimes caused by the production of fake voice recordings between 2013 and 2019 (Adam, 2022).

Forgery crimes in the document are regulated between Articles 204-212 of the TCK. If the nature of multimedia products produced with Deepfake technology is determined, it will be possible to determine which type of crime will be applied to the situation. In this regard, although Article 7 of the European Convention on Cybercrime, to which Turkey is also a party, regulates the issue of forgery crimes in electronic documents in parallel with forgery crimes in physical documents, there is no clear provision in our legislation on this issue. Technology giants such as Google, Meta, and Microsoft are in great competition in the field of developing Deepfake detection applications, and they also encourage the development of this field by sharing the source codes of the relevant projects. Another point that should be noted is that these detection applications are directly connected to Deepfake technology. This means that detection applications are basically based on the vulnerabilities of Deepfake technology. In this respect, when a detection application is developed, Deepfake software developers quickly close these vulnerabilities and disable the detection applications (Aitamurto, et al. 2022).

Fingerprints, color inconsistencies, texture distortions, optical flow analysis and the physical properties of the camera are the most important points of comparison in the detection process. In some detection techniques, the direct comparison criterion is anomalies in human head movements, eye movements and similar bodily impulses, and since the detection mechanisms are based on weaknesses in the general structure of the technology rather than technological gaps, it is almost impossible for Deepfake developers to close these gaps. Artificial neural networks in detection applications can be developed to work on these parameters, and it is not possible to talk about the inadequacy of detection applications in this respect.

The detection of these issues in the digital forensics process is quite costly and such a system is not currently established in our country. In this respect, it is an enigma how expert and digital forensics reports will emerge in the event that Deepfake products are subject to Turkish jurisdiction.

Analytic Discussion

Deepfake, which can be defined as artificial intelligence-supported disinformation, poses risks to many areas of life. The most frequently mentioned of these are those directly targeting the political sphere. The use of these technologies by authoritarian leaders in elections has begun to be called “deepfake democracy”, as in the 2024 Indian elections, and claims that global technology platforms have taken precautions in this regard are also false. A study discussing the harmful effects of fake videos on journalism and democracy argues that these videos not only damage the reputation of the person concerned, as seen in political candidates, but also cause confusion about what is real and what is not, increasing distrust of social and political actors in general. Indeed, political risk is not limited to elections. It examines how deepfake technology is weaponized by authoritarian

governments in their attempts to monitor, criminalize, and harass civil society, and how ongoing online and offline gender-based attacks on journalists, human rights defenders, and civil society leaders are being escalated, and how technology is being used to harm human rights, etc. Sam Gregory (2022: 713), who works both academically and practically in using it for positive purposes, explains it in detail. In the UK, he emphasizes ethical concerns about issues such as left-leaning media, deepfake, autonomous vehicles and weapons, privacy, facial recognition, and algorithmic bias and discrimination, as well as concerns about job loss due to automation. Due to the multifaceted nature of deepfake, there are different approaches in research on this subject. These differences also stem from the discipline studied, theoretical premises, and research methodology. In the literature, in addition to the positive uses of deepfake such as artistic creativity and satire; many risks have been underlined in terms of morality, ethics, epistemology of truth, and citizenship, human rights, and opposition to women's rights. In this article, I will briefly evaluate these discussions in the literature and highlight my preferred perspective. At the same time, I will try to underline critical conceptualizations and solution proposals by listing the discourses used in the fear of regulation or the tendency to downplay the threats that arise due to the uncertain structure of deepfakes (“deepfakes are not that many”, “not that effective”, “not new, they already existed”, “not just negative”, “not a matter of the present, but maybe in the future”, “after all, the audience will know”). As mentioned in the introduction, deepfakes are difficult to grasp simply because of their multifaceted character, their permeation into various communication processes, and the intertwining of production and content, distribution and reception processes. *Gray, J., Gerlitz, C., & Bounegru, L. (2018)* discusses deepfake processes under two headings: remix and manipulation: While deepfakes can be associated with art, creativity, education, satire and entertainment in terms of remix, they are also used for manipulation, as seen in disinformation and non-consensual porn. A collective report by *Jacobsen, B. N., & Simpson, J. (2023)*, analyzing more than seventy deepfake examples, shows that while some are examples of satire, art, or activism, others use comedy to glorify the powerful and attack marginalized communities. From the report, we learn that after they spread and harm people, they defend themselves with satire, as seen in the phrase “just kidding!”

The rapid development of deepfake technology and the inability of social media platforms to manage this technology have created a serious threat to the reliability of information systems. The possibility of manipulating social consciousness and political decisions with these technologies poses fundamental risks to democracy. Spiritual media is no longer at the center of the information flow, on the contrary, social behavior is formalized through the images, sounds and texts spread through social media.

This situation shows that platform governance is not just a technical matter; On the contrary, it is necessary to evaluate it in the context of ethics, political and social justice. Although deepfake technology was previously used for recreational and experimental purposes, it has now been turned into a weapon against journalists, political leaders and activists. Cases of opponents being ousted from influence through this technology have been observed in African countries. For example, video-interviewees on sax broadcast in Nigeria have had a great impact on local selections.

Platforms remaining either slow or neutral in such situations encourages the spread of false information. Although “Community Notes” and other user-based systems are presented as democratic responses to disinformation, the risk of manipulation is very high. An example of this is the systematic use of “discredited” labels against information published by feminist activists in Indonesia.

Some platforms use the method of deleting the content or reducing its visibility (shadow banning). Although this is not taken into account in terms of technical knowledge, in practice it limits the availability of alternative ideas in the social field. Basing algorithmic visibility priorities on corporate interests puts social responsibility in the background. In such circumstances, fact examiners' voices either become muffled or they react late.

At the same time, fact-checking partnerships are being established on many platforms, mostly in response to political pressure. Meta's reliance on fact-checking supporters in some countries has created a huge gap in the local and regional media ecosystems. This enables the public to accept false or even manipulated visual materials as truth.

It draws attention to areas where deepfake technology intersects with human rights. The creation and sharing of pornographic images named after female actors in the world shows that this technology has been turned into a gender-based means of oppression. Such situations have serious consequences not only in legal terms, but also in terms of creating social danger and trauma.

In some cases, simpler manipulations called "cheapfakes" also have a similar effect. In Pakistan and Sri Lanka, the statements of local politicians taken out of context had a serious impact on the behavior of the voters. This situation proves that much more than the technical accuracy of the image, it depends on its propagation speed and algorithmic support.

However, the problem of reliability increases for reporters and fact-checkers. Colleagues of a well-known fact-checking institution in Columbia were threatened and discredited by the public through a deepfake. These situations weaken the durability of the newspaper and the free press (Harvi, 2024).

From an analytical perspective, fact-checking processes are methods that require time and resources, but have long-term effects. However, platforms give more short-term reactions. For this reason, there is a coordination problem between platform policies and commercial fact-checking activities.

The main mechanism that formalizes the information behavior of social media users is algorithmic recommendation systems. These systems are not based on the user's past behavior, but sometimes on the platform's commercial interests. It has now been proven in many empirical studies that algorithms create ideological imbalance.

With the globalization process, the transmission of deepfakes has also become easier. The social discourse created in one country can influence the social discourse in another. For example, the publication of a deepfake video created in India before the election in Bangladesh shows the relevance of the interregional information environment.

It is not enough for the platforms to regulate themselves; Social institutions, legal defenders and the media must act together.

There are two main approaches to combating misinformation on social media platforms: user-based community models and systems based on professional fact-checking organizations. While each of these two models has certain advantages, their effectiveness and fairness need to be seriously debated (Borenstein, Warren, Elliott, Augenstein, 2025).

User-based systems such as Community Notes are important in terms of how platforms encourage user participation. In this model, people can add notes, explanations and corrections to the false information they see. This is especially widely used on the X (formerly Twitter) platform. This approach, which allows for quick reactions, seems important in the era of fast information.

However, the main drawback of this system is its inability to ensure impartiality. For example, during local elections in India, activists of a religious group systematically marked the posts of opponents by reducing their visibility. These manipulation attempts have damaged the democratic appearance of Community Notes. The openness of the system also makes it vulnerable to coordinated abuse.

Professional fact-checking, on the other hand, is carried out by specially trained journalists, researchers and experts. They investigate sources of information, assess context, and share their findings with the public. Reputable organizations such as Africa Check, Dubawa, Alt News, and Chequeado are all doing this in their regions.

The main advantage of this professional approach is its methodological reliability and accountability. Each fact analysis is presented with an explanation and the source is indicated. This transparency increases the credibility of the audience. However, the downside of this approach is that it takes time and cannot reach a large audience quickly. It is difficult for these systems to compete with the speed of social media.

A study conducted in Nigeria found that a post edited by the Community Notes system reached 500,000 people within 3 hours, while the Africa Check correction for the same post was distributed 48 hours later and reached only 50,000 people. While these figures show the advantage of speed on

platforms over professional structures, differences in quality and reliability remain significant (Chuai, Pilarski, Renault, Restrepo-Amariles, Troussel-Clément, Lenzini, Pröllochs, 2024).

In Indonesia, a feminist group's post was flagged as "contradictory" by Community Notes, resulting in 80% fewer shares. However, 48 hours later, a professional fact-checking agency declared the post to be completely accurate. This shows that incorrect user responses can cause public harm.

Furthermore, the fact that Community Notes only operates in the dominant language in many countries weakens its impact on disinformation in local languages. For example, in Tanzania, Community Notes was ineffective against misinformation in Swahili, but a local fact-checking group filled the gap.

In contrast, professional fact-checkers operate within a more normative framework. Their activities are regulated by specific codes (such as the IFCN principles). Community Notes does not have such mechanisms and operates without any accountability mechanisms.

Community Notes allows users to actively participate, but the consequences can be dangerous if the quality and intent of this participation are not measured. Professional structures, on the other hand, provide services with less participation but with higher quality.

Research shows that most audiences do not fully trust Community Notes. Adding different explanations to the same post can sometimes be confusing. However, professional fact-checkers are more trusted by the audience because they provide a unanimous and source-based response.

It can be concluded that these two models should work in a complementary way. Tools like Community Notes can be an initial intervention mechanism for a quick response, but professional fact-checking is essential for the final decision and public confidence.

In India, Alt News has established a collaboration mechanism with Community Notes, creating a model of synchronous work with a team of journalists who quickly check any user-added note. This model is a good example of synergy.

Critical Evaluation / Implications

Relying on technological tools to regulate platform governance is not enough to ensure global information security. While deepfake technology and algorithmic manipulations have fundamentally changed the nature of social media, the response of platforms has been either delayed or inadequate. This situation is not only a technical issue, but also results in damage to public trust, legal loopholes and violations of the principle of justice.

From an ethical perspective, the use of deepfake technology poses serious threats to personal freedoms and privacy. The public humiliation of women's rights defenders in Algeria through deepfake proves that this technology can be transformed into gender-based violence. Such cases not only prevent victims from contacting human rights defenders, but also create a deterrent effect on future public participation.

Platforms' responses are often based on commercial interests. This leads to a lack of a deep sense of responsibility. Although many platforms change their algorithms in response to criticism, the nature and effectiveness of these changes are not publicly disclosed. This lack of transparency undermines trust in the information environment.

Ethical analysis of algorithmic systems shows that systems that track and analyze user behavior are primarily profit-oriented. This is aimed at increasing the duration of use rather than the quality of information. Since sensational content such as deepfakes attracts more attention, platforms sometimes deliberately delay the dissemination of this content.

The uneven application of platform governance is more clearly observed in countries of the Global South. For example, in Tanzania, there are no Community Notes or fact-checking mechanisms for disinformation spread in Swahili. This creates information injustice and reinforces social inequality (Gosse, Burkell, 2020).

The weakness of legal frameworks is also ineffective in regulating the behavior of platforms. In many countries, legal regulation on deepfakes is either non-existent or poorly implemented. Despite the fact that a fake audio file affected the election in Nigeria, no criminal investigation was conducted on this issue. This situation shows that technology is moving faster than the legal system.

Platform governance cannot be limited to simply removing content or "adding labels." Ethical responsibility should encompass questions such as how user data is collected, what criteria are used to remove content, and who is made invisible.

A key issue here is the lack of "algorithmic justice" built into the platform. Algorithmic systems are often opaque, and users do not understand the logic of their decisions. This is contrary to the principle of accountability. Also, since most algorithms are designed with Western-centric data, they produce biased results in local contexts.

Another critical issue is who defines "dangerous content" and in what context. In Ethiopia, a statement by an opposition politician was removed by the system as a "radical call," but local experts interpreted the statement as "public protection." This highlights the dangers of centralized platforms regulating without understanding the local context.

Professional fact-checking actors also have their flaws. In some cases, these institutions operate under the influence of international financial donors and fail to adequately respond to local political realities. This can reduce their credibility and weaken their effectiveness in combating misinformation.

One of the dangerous situations is the emergence of a reflex among users to "approach all videos with suspicion". This skepticism leads to a decrease in public participation and a disregard for political decision-making. In other words, deepfake technology not only creates fake information, but also questions what is true. This "epistemic injury" undermines social capital.

In addition to those who argue that regulation may be necessary to prevent the harms of deepfake culture without causing other harms, and that regulating the algorithmic circulation of messages shared for corporate profit is not an attack on censorship or freedom of expression, suggestions such as building democratic resistance and creating public pressure are also being developed by pointing out the drawbacks of regulation (Dan, 2021).

What is very important here is this: Despite its technical dimensions, the issue is actually a socio-cultural issue and therefore difficult to solve with technical methods. Jacquelyn Burkell and (*Matthews, Kidd, 2024)* also emphasize the socio-cultural and material nature of deepfake videos, and in their studies they mention that their sophisticated technology and metaphysical nature, both real and unreal, have become invulnerable to many technical, legal and regulatory solutions, and that it is similarly difficult to define the harm they do to the targeted individuals.

There are those who point out the danger that if algorithmic moderation is adopted as a solution, the pro-democracy discourses of marginalized people and activists may also be censored. Accordingly, algorithmic moderation currently cannot reliably detect hate speech: Therefore, it is recommended that the moderation practice be evaluated very carefully, as its democratic harms may outweigh its democratic benefits, and instead of moderation, platforms are advised to consider new algorithmic moderation in line with anti-racist, feminist and democratic theories. Those who emphasize that anti-deepfake measures can harm creative industries, such as documentary production, also point to the need to develop a media literacy agenda and raise public awareness on this issue. Here, listening to those who say that we have a responsibility to draw a line between experimental, expressive uses and sexual, political or other forms of visual and auditory manipulation and exploitation, it can also be considered that critical artificial intelligence (AI) and data literacy can be one of the solutions. In the roundtable meeting on "Verification and Artificial Intelligence" held in 2023, one of the various suggestions developed by representatives of leading verification platforms in Turkey, global technology companies and academics is literacy. There are also attempts to conceptualize media, information and digital literacy more broadly and to link it to critical AI and data literacy practices for citizen participation (Matthews, Kidd, 2024). Such literacy, as Aristea Fotopoulou (2021) has also stated, is not a technical literacy but must necessarily be associated with themes of inequality and justice: Work on this axis should not be limited to journalists and academics, but should be directed towards civil society, and critical data literacy should be developed in a way that provides agency, contextual awareness and social responsibility. Jonathan (McCosker, 2022) have also argued in their work on data infrastructure literacy that new forms of mobilization, intervention and activism are being opened up. Those who argue that awareness, mechanisms and

ethical paths should be created to combat data-based discrimination propose data justice in connection with a social justice agenda against inequalities, discrimination and the exclusion of certain groups.

When we look at what has already been done, for example, in the handbook prepared by UNESCO for journalism educators on reporting on artificial intelligence (2023), the deepfake issue is specifically addressed, underlining its harm to democratic culture and the need to combat it. Although the issue of discrimination and justice is mentioned in the ethical guidelines published by the European Commission (2019) for reliable artificial intelligence, and although there are ethical guidelines being developed by press professional organizations in various countries, it is still difficult to say that justice-based approaches are gaining ground. There is a need for studies that adopt a more rights-oriented and justice perspective, examples of which have started to be seen in recent years, although few in number. An example of materials produced for educational studies to be carried out from such a perspective is the <A+> Alliance (2020). This global, multidisciplinary, feminist group, which carries out future-oriented studies to improve gender equality with the help of technology and innovation, consists of academics, activists and technologists. It provides a human rights and gender equality toolkit for educators and students, modules examining equality, prejudice, intersectionality, and workshop materials. The study, which is an example of good practice in itself, can, in our opinion, pave the way for the development of other good practices (Shade, 2023).

Finally, in this article, I have tried to briefly evaluate the discussions on the subject in the literature and to highlight justice and rights-oriented approaches. At the same time, I have tried to underline critical conceptualizations and solution proposals against the discourses that emerge due to the ambiguous structure of deepfake and against the attempts to reduce the subject to a technical issue. Although images and videos are never “objective” and reality is constructed within networks of power relations, I tend to believe that reaching the truth should remain an ideal in the post-truth world. At this point, we do not have to choose between political and social use: They are all important, interrelated, and urgent. From this point on, we should be prepared for the potential use of deepfake videos in politically manipulating mechanisms such as elections, limiting our limited democratic rights even further, and increasing and deepening existing inequalities in society. Deepfake videos, especially those with sexual content and pornography, will never eliminate the harm they have done to the women they target, even if they are later removed. At this point, the concepts of media literacy and especially "deepfake literacy" come to the fore.UNESCO recommends that educators not only question students' "believing what they see" reflex, but also provide technical analysis skills. It is particularly recommended to analyze clues such as facial synchronization, tone of voice, and light-shadow inconsistencies in videos.

UNESCO's guidelines also state that media literacy in education should not be limited to the classroom, but should be integrated into journalism, citizenship and human rights courses. In its document titled "Guidelines for Teaching Journalism with AI", published in 2023, UNESCO examines in depth the effects of artificial intelligence (AI) on media and journalism. This document is a guideline on the use of AI technologies in journalism teaching. It is emphasized that new approaches are needed in the educational process, especially in connection with the increasing impact of deepfake technologies. UNESCO recommends that students be educated not only as information consumers, but also as critical thinking individuals with technical analysis skills. Not immediately assuming what they see is true and developing visual-analytical skills are shown as one of the main goals. It is recommended to pay attention to technical indicators such as facial expressions, tone of voice, and light-shadow discrepancy to distinguish deepfake videos. This guideline does not limit media literacy only to journalism lessons. At the same time, it recommends that it be integrated into citizenship, human rights and technology subjects (Hight, 2022).

A+ Alliance is an international initiative that aims to integrate gender equality with technology. The toolkit they published in 2020 offers educators the opportunity to question the gender-based effects of artificial intelligence and develop lessons in this area. A+ Alliance is an innovative platform that emphasizes that gender equality in the age of digital transformation is not only a social but also a technological issue. The organization focuses on the idea that technology is not neutral, but can reproduce social prejudices. Artificial intelligence systems and algorithms are often trained with data

that reinforces gender stereotypes. For this reason, A+ Alliance advocates the active participation of women and minority groups in the design, development and implementation processes of technology. The organization states that digital technologies should not only produce technical solutions, but also observe social justice. Algorithmic justice is one of the most emphasized concepts in this context. A+ Alliance provides many examples showing that algorithms carrying gender biases can affect decision-making processes without being noticed. Recruitment software, credit scoring systems and facial recognition technologies are especially vulnerable to these biases. In these technologies, gender discrimination can affect not only individuals but also all social structures. In this context, A+ Alliance emphasizes the concept of "ethical artificial intelligence". Ethical design principles, data diversity and participatory innovation are the cornerstones of this model. The organization's work aims to design artificial intelligence systems in a transparent and accountable way. At the same time, it argues that users should have the knowledge and tools to question these systems.

MediaSmarts, based in Canada, is an NGO that provides digital media literacy training, especially for children and youth. Its program called "Break the Fake" aims to provide skills to recognize, analyze and verify deepfakes and other fake content.

Program content:

• Verification of video content with reverse image search
• Analysis of visual and audio cues (eye blinking, voice synchronization, etc.)
• Classroom applications analyzing social media content

The program is designed for different age groups from primary school to high school.

The "Break the Fake" program prepared by MediaSmarts is a systematic and educational response to the increasing problem of false information in the digital media environment. This initiative encourages people – especially young users – to critically analyze the information they encounter on the Internet. The main purpose of the program is to provide users with practical and easily applied methods to evaluate the accuracy of information. “Break the Fake” is based on four main steps: finding the source, examining the source, comparing the information with other sources and finding evidence. This seemingly simple, yet effective approach formalizes the basic skills of numerical warfare. The program provides not only theoretical but also practical tools. Through lesson plans, workshops and interactive tests, teachers and students can practice the structure of the program. The famous video called “House Hippo 2.0” shows how convincingly visual information can be presented to the viewer. This video explains the importance of focusing criticism not only on texts but also on figures and videos. The main power of the program is that it turns its users from passive information acceptors to active researchers. Students are taught not only to "examine facts" but also to create their own "research mechanisms". For example, within the framework of the program, students formalize their self-verification strategies and use individual search methods. This, in turn, strengthens the civilization of independent decision-making and inquiry. MediaSmarts considers it important not only to have technical knowledge in order to protect itself from the demands of technology, but also to develop a sense of moral responsibility. As executive director of WITNESS, Sam Gregory guides educational programs aimed at re-establishing media warfare and legal defense against the backdrop of the spread of artificial intelligence and deepfake technologies. He considers it important to investigate in parallel the positive and negative effects of technology that can be used in the field of human rights. The "Prepare, Don't Panic" program, which was implemented with the initiative of Gregory, aims to ensure that users and human rights defenders are ready against media manipulation. This program is equipped with training modules that explain numerical manipulation forms, especially deepfake technology. Gregory prioritizes not only technical training but also strengthening ethical reflexes. During the training, users are taught how to interpret the images created by artificial intelligence (Dan, 2021).

At the same time, metadata analysis methods are applied to ensure the reliability of video and photographic evidence. In WITNESS programs, information is given about both technical tools and legal frameworks. The purpose of the programs, in addition to protecting individuals, is also to deter society from being exposed to information threats. Gregory's approach is based on the principle of "being prepared and thinking analytically" instead of panic and technophobia. Educational programs

are considered for video journalists, citizen journalists, legal advocates and teachers. In these programs, technical and ethical approaches against the manipulation of "evidence based on testimony" are taught. Gregory considers not only the dissemination of information important, but also the way in which it is explained. In the studies, the responsible sharing and archiving of images that violate human rights is also emphasized by some authorities. In WITNESS's educational programs, regionally and linguistically localized educational materials are presented. This makes it easier for the program to reach global audiences. Sam Gregory is the director of the international NGO called WITNESS and his project "Fortify the Truth" addresses the effects of deepfake and artificial intelligence content on human rights.

According to Gregory:

• Deepfake technology endangers the concept of democratic witnessing.

• The goal of education is not only technical analysis, but also awareness of social effects.

• In schools, students should be taught the skills to both question content and discuss ethical values.

Gregory sees deepfake literacy not only as “technical filtering” but also as an ethical, social and political skill.

Aristea Fotopoulou is a Reader (Associate Professor) in Digital Communication, Culture and Society at the University of Brighton. Her research focuses on the intersection of digital media, data-based technologies and social justice. She has made significant contributions to feminist science and technology studies (STS), queer theory and digital activism.

Fotopoulou’s work addresses the impact of digital culture on gender, identity and power relations. In her book “Feminist Activism and Digital Networks: Between Empowerment and Vulnerability”, she analyses the empowering and fragile effects of digital networks on feminist activism. In her work “Feminist Data Studies: Big Data, Critique and Social Justice”, she also highlights the importance of feminist critique and social justice in the age of big data.

Her research examines the social and cultural dimensions of the "Quantified Self" movement by examining the practices of data sharing and self-tracking in individuals' daily lives. In this context, she evaluates the effects of wearable technologies and health applications on individuals' body perception and health behaviors.

Fotopoulou was the Chair of the Digital Culture and Communication Department of the European Communication Research and Education Association (ECREA) between 2016-2018. She also aimed to develop creative solutions for health and well-being issues through art and data as the Principal Researcher of the "ART/DATA/HEALTH" project.

She began her academic career with a PhD in Media and Cultural Studies at the University of Sussex, and then did postdoctoral research at Goldsmiths, University of London and University of Sussex. She also served as a visiting researcher at the Science and Justice Research Center at the University of California, Santa Cruz.

Ethical criticism should also be applied to journalism. Some news portals deliberately publish deepfake or manipulated images for the purpose of “clickbait”. In this case, the problem is not only in technology, but also in the decline of media professionalism.

If these processes are not prevented, a “post-information era” may arise in the long term, in which the public does not trust any information. This is a threat not only to democracy, but also to social stability.

Conclusion

While education-based media literacy approaches provide a layer of security at this point, they are not sufficient on their own. It is imperative that platform companies develop more transparent, inclusive, and ethically based governance policies. At the same time, fact-checking mechanisms need to be restructured in accordance with the principles of independence and professionalism. In conclusion, establishing the balance between freedom of expression and combating disinformation in the age of deepfakes and synthetic media is possible not only with technical solutions, but also with

a multi-layered approach that integrates ethical, social, and political perspectives. In this context, considering media literacy programs and platform governance policies together and supporting them with global good examples will increase the resilience capacity of democratic societies. This study dealt with the effects of deepfake technology and artificial intelligence-supported media manipulation on the digital information environment, social structure and platform management systems with a multifaceted perspective. The obtained findings clearly revealed that the digital media ecosystem is shaped not only by technological developments, but also by ethical, cultural and political structures.

The examples and literature data evaluated during the research showed that the policies adopted by the platforms in content management are often far from transparency and shaped by algorithmic biases. Participatory models, especially community-based moderation systems, have been shown to be vulnerable to organized disinformation campaigns as much as they offer potential opportunities. At the same time, despite the methodological strength of professional fact-checking structures, the difficulties experienced in quick access to the masses create a double impasse in the information war: the dilemma of speed and reliability.

However, shaping the behavior of platforms only with market dynamics damages users' rights and society's trust in information. The rapid spread of content created by deepfake technology through social media threatens not only the reputation of individuals, but also public safety, election processes and democratic culture.

One of the main findings of the study is that platform governance should be supported not only by technical interventions, but by a multi-layered and multi-actor system. In this context, algorithmic transparency should be ensured, user participation should be monitored and professional accuracy mechanisms should be independently strengthened. In addition, the expansion of media literacy training and the creation of ethical awareness in the area of digital justice will play a vital role in creating social resistance points in the long term.

References

1. A+ Alliance. (2020). Affirmative action for algorithms: Artificial intelligence, machine learning, & gender.
2. Aitamurto, T., et al. (2022). Examining augmented reality in journalism: Presence, knowledge gain, and perceived visual authenticity. New Media & Society, 24(6), 1281–1302.
3. Burkell, J., & Gosse, C. (2019). Nothing new here: Emphasizing the social and cultural context of deepfakes. First Monday.
4. Cover, R. (2022). Deepfake culture: The emergence of audio-video deception as an object of social anxiety and regulation. Continuum.
5. Dan, V. (2021). Fake videos: Challenges for journalism and democracy. Journalism & Mass Communication Quarterly, 98(3), 643–645.
6. Garg, A., & Kumar, A. (2024). Artificial Intelligence and Political Deepfakes: Shaping Citizen Perceptions and Democratic Discourse. *Journal of Creative Communications, 19*(1), 45–60.
7. Harvi, P. L. K. (2024). Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media. *ResearchGate*.
8. UNESCO. (2024). User Empowerment Through Media and Information Literacy to Counter the Evolution of Generative Artificial Intelligence. *UNESCO Policy Brief*.
9. Gao, Y., Zhang, M. M., & Rui, H. (2024). Can Crowdchecking Curb Misinformation? Evidence from Community Notes. *Gies College of Business, University of Illinois Urbana-Champaign*.
10. Borenstein, N., Warren, G., Elliott, D., & Augenstein, I. (2025). Can Community Notes Replace Professional Fact-Checkers? *arXiv Preprint*.
11. Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., & Pröllochs, N. (2024). Community-Based Fact-Checking Reduces the Spread of Misleading Posts on Social Media. *arXiv Preprint*.
12. Gosse, C., & Burkell, J. (2020). Politics and porn: How news media characterize problems presented by deepfakes. Critical Studies in Media Communication, 37(5), 497–511.

13. Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. Big Data & Society, 1–13.
14. Hight, C. (2022). Deepfakes and documentary practice in an age of misinformation. Continuum, 36(3), 393–410.
15. Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. Information, Communication & Society.
16. Matthews, T., & Kidd, I. J. (2024). The ethics and epistemology of deepfakes. In C. Fox & J. Saunders (Eds.), The Routledge Handbook of Philosophy and Media Ethics. Routledge.
17. McCosker, A. (2022). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. New Media & Society, 1–18.
18. Rini, R. (2020). Deepfakes and the epistemic backstop. Philosopher's Imprint, 20(24).
19. Shade, L. R. (2023). From media reform to data justice: Situating women's rights as human rights. In M. Gallagher & A. V. Montiel (Eds.), The Handbook of Gender, Communication, and Women's Human Rights. Wiley.
20. Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. Big Data & Society, 1–14.
21. Watson, L. (2021). Epistemic rights and why we need them. London: Routledge.
22. WITNESS. (n.d.a). Deepfakes, synthetic media and generative AI.

Received: 10.02.2025
Accepted: 31.05.2025